Automated Pedestrian Recognition Based on Deep Convolutional Neural Networks

Obaida M. Al-Hazaimeh and Ma'moun Al-Smadi

***Abstract*—Accurate and precise pedestrian detection play a major role in analyzing and understanding its behavior. Pedestrian recognition is a challenging task due to body deformation, weather, and lighting conditions variations. Various techniques combine feature extraction with support vector machine. However, deep Convolutional Neural Networks (i.e. CNNs) achieved promising results in various recognition tasks. Although, CNNs require large databases and expensive computations, it can outperform other algorithms more accurately. In this paper, we proposed a deep learning technique for automatic pedestrian recognition based on image normalization and CNN architecture. The proposed architecture learns pedestrian representation adaptively to achieve efficient recognition with higher accuracy and lower pre-processing time. The experimental results show that the proposed technique out performs conventional methods superiorly.**

***Index Terms*—Neural network, machine learning, pedestrian recognition, deep learning.**

I. Introduction

Object detection, tracking and classification have many applications in surveillance and monitoring systems. In intelligent transportation systems (i.e., ITS) object detection can be applied for pedestrian recognition, vehicle classification, traffic sign and lane detection. Pedestrian recognition has been one of the core research areas in traffic surveillance for a long period of time, since it provides valuable information for event detection and behavior understanding [1]. Image processing and computer vision are widely applied in detection and recognition applications. The quality of image or video together with human body deformation, lighting and weather conditions are the main challenges in pedestrian recognition. Pedestrian image may vary in shape, color, pose and orientation according to its motion and behavior. Moreover, the presence of multiple pedestrian in a single scene may raise the challenge of occlusion and real time processing [1], [2].

Many research activities focus on pedestrian detection and recognition. Most of them utilize either motion segmentation combined with template matching or descriptive feature extraction followed by machine learning like support vector machine [3], [4]. Pedestrian recognition using color histogram matching was discussed in [5], by extracting color histograms from three horizontal partitions of the human image. Color features was used to model human body over the principle axis in [6], in which locating the principal axis was affected by cluttered background and crowded scenes. In [6], deformable part-based detectors were investigated. Many researchers explore the variation in part modelling and deformation, while major efforts were focused on discriminative low-level features.

Discriminative feature extraction increases object diversity and improve discrimination and recognition quality. Thus, pedestrian detection and recognition with richer and higher dimensional representations become easier and provide better results. A large set of features have been discussed in the literature (i.e., edge, color, texture, local shape features, and covariance features) [7]-[15]. The improvement in recognition performance depends on the growth in feature diversity. Recently, a variety of discriminative features have been used in pedestrian detection like Scale Invariant Feature Transformation (i.e., SIFT), speeded up Robust Features (i.e., SURF), Histogram of Oriented Gradient (i.e., HOG) and Haar-like features [16], [17]. Each of them has its own limitations and drawbacks. Advanced techniques like deep convolutional neural networks (i.e., DCNN) form a hot area of research for pedestrian detection and recognition that have a promising future.

In this paper, we have proposed a pedestrian detection and recognition technique that distinguish pedestrian from other types of road users such as car, bus, and bicycle. The proposed technique takes advantage of recent progresses in deep convolutional neural networks (i.e., DCNN) that can learn unique pedestrian features quickly and accurately. Images from data set are resized into fixed scale, normalized by zero-centered mean with unity standard deviation and feed into the proposed DCNN to identify pedestrian from non-pedestrian.

Generally, the proposed technique has five linear convolutional layers and each one is followed by Rectified Non-Linear (i.e., ReLU) layers. The max pooling layers follow all convolutional layers except the third layer. Sigmoid activation function at the last layer of the cascaded feature extraction.

This paper will be arranged as follows. Section II presents Deep Convolutional Neural Network and the proposed deep learning architecture. Next, the experimental results of the proposed technique with the analysis and discussion are presented in section III. Finally, the conclusion is presented in Section IV.

II. Deep Convolution a Network

Convolutional Neural Network (i.e., CNN) is a supervised machine learning technique that is inspired by the human brain and neurons. In 1989 and 1990 Yann LeCun *et al*., [8]

Manuscript received July 3, 2019; revised August 20, 2019.

The authors are with the Al-Balqa'Applied University, Jordan (e-mail: drobaidam@yahoo.com, masmadi@bau.edu.jo).

doi: 10.18178/ijmlc.2019.9.5.855

introduced several aspects of modern CNNs, such as feature maps, sub-sampling, and shared weights. Since then several modifications and improvements have been proposed in the literature at the expense of computational complexity that require special hardware.

Deep learning models achieve non-linearity using an activation function. It is applied after each convolutional layer. Many CNN architectures use rectified linear unit (ReLU):

$$f(x) = \max(0, x) \tag{1}$$

Other CNN architectures may employ sigmoid or hyperbolic tangent as an activation function as shown in equations 2, and 3 respectively:

$$f(x) = \sigma(x) = \frac{1}{1+e^{-x}} \tag{2}$$

$$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1 \tag{3}$$

Convolutional layer performs feature map extraction by convolving a set of kernels with the layer input using specific kernel function. For each layer l, the input x_i to the ith feature map is convolved with a kernel $w_{ij}^l$of a specific size and substituted in the activation function to get the output as:

$$x_i^l = f\left(\sum_j w_{ij}^l * x_j^{l-1}\right) \tag{4}$$

In this paper, gray scale images are used, thus the inputs and the feature map are both two dimensional with size of $W\times H$. The hyperparameters of a convolutional layer are the number of kernels, stride and kernel sizes. Each input image of size $W\times H$ is convolved with N kernels of size $W_k \times H_k$ and a stride S will produce an output feature map of size $W_o \times H_o$ where $W_o = (W - W_k)/S + 1$ and $H_o = (H - H_k)/S + 1$.

The output of the previous step may contain repetition or redundant information. Thus, a pooling layer is used to eliminate redundant information from the feature maps. Usually, max pooling average pooling or L2-norm pooling is used after convolutional layer. It is applied on an $n\times n$ regions with the feature map. Depending on the kernel and stride size S it can be either overlapping or not as shown in Fig. 1. To make it clear, pooling layers can help to reduce over fitting and exclude redundant since it reduces the number of parameters as well as the size of the feature maps.

2	1	1	8
3	5	7	4
6	1	3	1
2	3	1	4

5	8
6	4

5	7	8
6	7	7
6	3	4

Fig. 1. Example of 2×2 max pooling with a stride of 2 and stride of 1.

Fully-connected layers are used as final layers in CNNs. The last layer outputs the estimations. Multi-class networks use soft-max classifier in the final layer while binary classification is done using the sigmoid function. Same as regular neural networks, the computational units in a fully-connected layer are connected to every unit of the previous layer. Unlike convolutional and pooling layers fully-connected layers are one-dimensional.

III. Proposed Deep Learning Network Architecture

The main contribution of this paper is a deep learning technique for automatic pedestrian recognition based on image normalization and DCNN architecture that will enhance the object detection services (i.e., recognition accuracy). The flow diagram of the proposed technique can be described by Fig. 2.

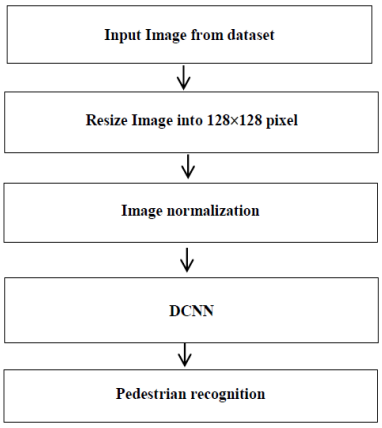


Fig. 2. Flow diagram of the proposed technique.


(a) Pedestrian


(b) Pedestrian


(c) Pedestrian


(d) Pedestrian


(e) Non- pedestrian


(f) Non- pedestrian


(g) Non- pedestrian


(h) Non- pedestrian

Fig. 3. Dataset.

To make it clear, the proposed technique is based on two of parallel and sequential steps, which are partially automated: Step 1, obtaining an appropriate dataset that contain pedestrian and non-pedestrian. Step 2, the selected images are resized, normalized and feed into the proposed DCNN to achieve pedestrian recognition.

The data set used in this paper was taken from MIO-TCD classification dataset [18], it contains 6156 pedestrian images from a total of 648,959 images of various categories and sizes. Images are colored and contain objects (pedestrian and Non-pedestrian) with different sizes, poses and orientations. A total of 3100 images were selected from the data set, 1550 Pedestrian and 1550 Non-pedestrian images as shown in Fig. 3.

CNN require an equally sized input images, thus all images are resized into 128×128 pixel. To make sure that the features are zero-centered and have a unity standard deviation, standardization method is applied to normalize images by subtracting the mean from each pixel and then dividing by standard deviation as:

$$I = \frac{x-\mu}{\sigma} \tag{5}$$

where, μ is the mean and σ is the standard deviation of each image. In other words, the mean and standard deviation are computed over the training set and the same values are used for normalizing the test set. Standardization ensures that the features are zero-centered and have a standard deviation of 1 [11], [18], [19].

The proposed DCNN architecture contains 4 convolutional layers and 2 fully-connected layers as shown in Fig. 4 and Fig. 5 respectively. After the first, second and forth convolution layers, contrast normalization, pooling, and nonlinear function are applied, while the output of the third convolution layers is directly feed into the forth layer.

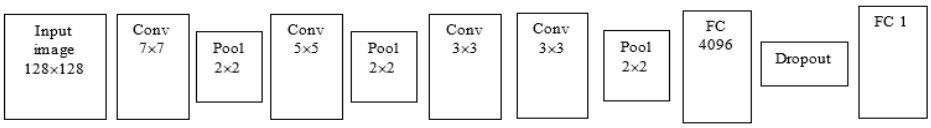


Fig. 4. Architecture of DCNN for pedestrian recognition (Conv = Convolution; Pool = Pooling; FC = Fully connected).

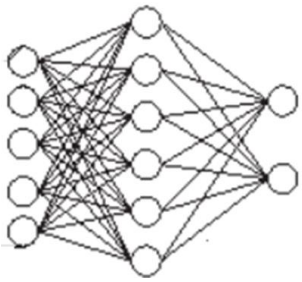


Fig. 5. Fully connected layers (FC1).

TABLE I: DEEP CNN ARCHITECTURE

	Kernal size	# of kernels	Stride Size	Output Size
Convolution 1	7×7	32	1	122×122×32
Max pool 1	2×2	-	2	61×61×32
Convolution 2	5×5	64	1	57×57×64
Max pool 2	2×2	-	2	28×28×64
Convolution 3	3×3	128	1	26×26×128
Convolution 4	3×3	256	1	24×24×256
Max pool 4	2×2	-	2	12×12×256
Fully connected				4096
Fully connected				2 Classes

The first convolutional layer applies 32 filters of size 7×7, followed by a max 2×2 pooling layer. The second convolutional layer has 64 filters with a size of 5×5, followed by a max 2×2 pooling layer. The third convolutional layer has 128 filters with a size of 5×5 followed by the forth convolutional layer has 256 filters with the same kernel size as previous layer. After that a 2×2 max pooling is applied again. The output of the convolution layer is reshaped as a feature vector and fed to the fully-connected layers. The output layer has a single unit in the classification task to predict whether image is pedestrian or not. The parameters of the described layers are illustrated in Table I.

IV. EXPERIMENT AND DISCUSSION

The proposed architecture is evaluated and compared on a standard benchmark dataset. Experiments were performed using MATLAB R2015a on DELL laptop (2.3 GHz, 4th generation core i5 CPU with 8-GB RAM, SSD storage and windows 10 64-bit).

For pedestrian recognition task, the network will distinguish between pedestrian and non-pedestrian. Two DCNN networks were trained and tested: one DCNN without normalization step and one with normalization. Both networks were trained for 20 epochs using Weka software at learning rate of 0.001. Cross-entropy loss function was used for binary classification. In order to evaluate the pedestrian recognition, the true positive rate (TPR), Recall and precision are defined as:

$$TPR = \frac{\text{Number of correctly recognized pedestrian}}{\text{Total Number of Pedestrian}} \tag{6}$$

$$TNR = \frac{\text{Number of incorrectly recognized pedestrian}}{\text{TotalNumberofnon}-\text{Pedestrian}} \tag{7}$$

$$\text{Presision} = \frac{\text{True positive}}{\text{True positive+False Positive}} \tag{8}$$

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive+False negative}} \tag{9}$$

The confusion matrixes for both networks are shown in Table II, and Table III. It is clear from the obtained results that the classification accuracy is improved by using image normalization. It achieves a true positive rate of 94.9% and true negative rate of 94.6%. while the true positive and negative rates without normalization were 92.6% and 91.1% respectively. Hence, image normalization improves the true positive rate by 2.3%, while true negative rate was improved by 3.5%, which indicate a more accurate discrimination.

TABLE II: CONFUSION MATRIX FOR DCNN WITHOUT IMAGE NORMALIZATION

		Predicted Class	
		Pedestrian	**Non-Pedestrian**
True Class	Pedestrian	1436 92.6%	114 7.4%
	Non-Pedestrian	138 8.9%	1412 91.1%

TABLE III: CONFUSION MATRIX FOR DCNN WITH IMAGE NORMALIZATION

		Predicted Class	
		Pedestrian	**Non-Pedestrian**
True Class	Pedestrian	1471 94.9%	79 5.1%
	Non-Pedestrian	83 5.4%	1467 94.6%

Precision and Recall result for both Networks are shown in Table IV below. Before image normalization the proposed DCNN architecture achieves 0.9123 precision and 0.9265recall. After adding image normalization, the precision increase to 0.9465 and the recall rise to 0.9490. The improvement in precision due to image normalization is 3.42%, which is higher than the improvement in recall that is only 2.43%.

TABLE IV: PRECISION AND RECALL FOR BOTH DCNN

	Precision	**Recall**
Without Normalization	0.9123	0.9265
With Normalization	0.9465	0.9490

To compare the results with the state of the art, HOG feature Descriptor combined with support vector machine classifier was used. The Confusion matrix for HOG descriptor using SVM is shown in Table V. The TPR is 90.6% and TNR is 88.9%.

TABLE V: CONFUSION MATRIX FOR HOG USING SVM

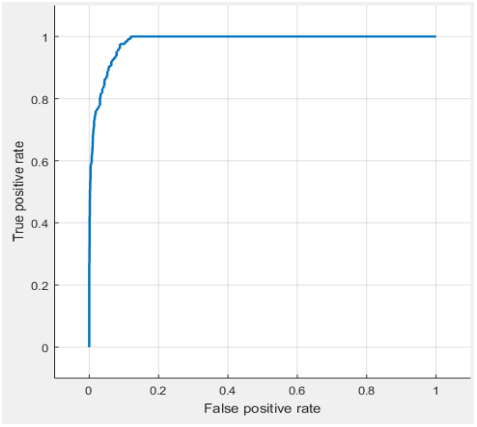
		Predicted Class	
		Pedestrian	**Non-Pedestrian**
True Class	Pedestrian	1405 90.6%	145 9.4%
	Non-Pedestrian	172 10.1%	1378 88.9%

Table VI compare the proposed DCNN architecture against HOG descriptor with SVM. The proposed architecture with image normalization achieves the best precision and recall of 0.9547 and 0.9490 respectively. While moderate precision and recall of 0.9123 and 0.9265 were achieved without image normalization as compared to the standard HOG descriptor with SVM which achieve the lowest precision and recall of 0.8909and 0.9065 respectively. Thus, the proposed technique provides better results with 2.14% higher Precision and 2% higher Recall without image normalization. On the other hand, image normalization raises the Precision and Recall against HOG descriptor with SVM by 6.38% and 4.25% respectively.

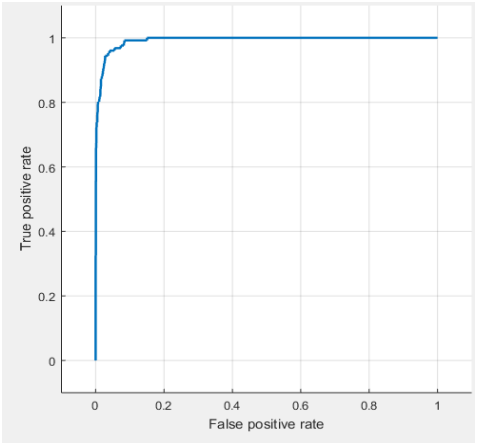
TABLE VI: PRECISION AND RECALL FOR BOTH DCNN AND HOG

	HOG+SVM	**DCNN**	**DCNN + Normalization**
Precision	0.8909	0.9123	0.9547
Recall	0.9065	0.9265	0.9490

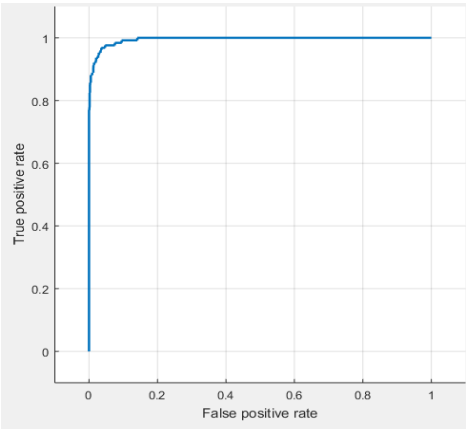
The Receiver Operating Characteristics curve (ROC curve) for pedestrian recognition using HOG descriptor with SVM, DCNN and DCNN with Normalization is shown in Fig. 6. The lower false positive rate (FPR) is of more interest, for example if FPR is less than 0.1 the true positive rate (TPR) will be around 0.9 for both HOG descriptor with SVM and DCNN without normalization. On the other hand, the true positive rate for DCNN with Normalization will be around than 0.95 for the same low false positive rate.



(a) HOG+SVM



(b) DCNN



(c) DCNN+Normalization

Fig. 6. ROC curve for pedestrian recognition using HOG+SVM, DCNN, and DCNN+Normalization.

The use of bicycles and motorcycles in the non-pedestrian dataset is the main reason for incorrect recognition in both pedestrian and non-pedestrian cases. This is due to the fact that a human body forms a major part of such objects. Thus,

incorrect pedestrian recognition occurs if the lower part of the pedestrian is not clear and recognized as a bike. On the other hand, if the bike body is not clear enough, the motorist (bike rider) can be misclassified as a pedestrian. Fig. 7 below shows some examples of image results for correct and incorrect pedestrian recognition.

(a) Correctly recognized as pedestrian (b) Incorrectly recognized as pedestrian

(c) Correctly recognized as pedestrian (d) Incorrectly recognized as pedestrian

(e) Correctly recognized as non-pedestrian (f) Incorrectly recognized as non-pedestrian

(g) Correctly recognized as non-pedestrian (h) Incorrectly recognized as non-pedestrian

Fig. 7. Examples of correct and incorrect recognition.

V. Conclusion

Pedestrian recognition in the images is a challenging work due to body deformation, weather, and lighting conditions variations. In this paper, we have proposed deep learning technique for automatic pedestrian recognition based on zero-centered image normalization and Deep Convolutional Neural Networks (CNN) architecture to improve the object recognition accuracy. In the proposed technique, the object is detected and then the detected object under different conditions can be accurately classified (i.e. Pedestrian, Non-Pedestrian). The proposed architecture learns pedestrian representation adaptively to achieve efficient recognition with better discrimination, higher accuracy and lower preprocessing time. The experimental results show that the proposed technique out performs conventional methods superiorly.

Acknowledgment

The authors would like to thank all the people who support this research work. Especially our colleagues from Al-Balqa' Applied University (BAU), Jordan.

References

[1] M. Al-Smadi, A. Khairi, and R. A. Salam, "Traffic surveillance: A review of vision based vehicle detection, recognition and tracking," *International Journal of Applied Engineering Research*, vol. 11, no. 1, pp. 713-726, 2016.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. pp. 886-893. [Online]. Available: https://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf

[3] M. Al-Smadi, A. Khairi, and R. A. Salam, *Cumulative Frame Differencing for Urban Vehicle Detection*, p. 100110G.

[4] O. M. Al-Hazaimeh, M. Al-Nawashi, and M. Saraee, "Geometrical-based approach for robust human image detection," *Multimedia Tools and Applications*, pp. 1-25, 2018.

[5] U. Park, A. K. Jain, I. Kitahara, K. Kogure, and N. Hagita, "Vise: Visual search engine using multiple networked cameras," in *Proc. 18th International Conference on Pattern Recognition*, pp. 1204-1207.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, 2010.

[7] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. 10th ECCV*, pp. 262-275.

[8] P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable deep network for pedestrian detection," in *Proc. CVPR*, pp. 899-906.

[9] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1030-1037.

[10] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. BMVC*, 2009.

[11] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE International Conference on Computer Vision*, pp. 32-39.

[12] A. D. Costea and S. Nedevschi, "Word channel based multiscale pedestrian detection without image resizing and using only one classifier," in *Proc. CVPR*, pp. 2393-2400.

[13] S. Paisitkriangkrai, C. Shen, and A. V. D. Hengel, "Efficient pedestrian detection by directly optimizing the partial area under the ROC curve," in *Proc. ICCV*, pp. 1057-1064.

[14] N. Gharaibeh, O. M. Al-Hazaimeh, B. Al-Naami, and K. M. Nahar, "An effective image processing method for detection of diabetic retinopathy diseases from retinal fundus images," *International Journal of Signal and Imaging Systems Engineering*, vol. 11, no. 4, pp. 206-216, 2018.

[15] M. Al-Nawashi, O. M. Al-Hazaimeh, and M. Saraee, "A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments," *Neural Computing and Applications*, vol. 28, no. 1, pp. 565-572, 2017.

[16] L.-C. Chen, J.-W. Hsieh, H.-F. Chiang, and T.-H. Tsai, "Real-time vehicle color identification using symmetrical SURFs and chromatic strength," in *Proc. International Symposium on Circuits and Systems*, pp. 2804-2807.

[17] A. Ma'moun, O. M. Al-hazaimeh, N. Alhindawi, and S. M. Hayajneh, "A dual curvature shell phased array simulation for delivery of high intensity focused ultrasound," *Computer and Information Science*, vol. 7, no. 3, p. 49, 2014.

[18] Z. Luo, B. Frederic, C. Lemaire, J. Konrad, S. Li, A. Mishra, A. Achkar, J. Eichel, and P.-M. Jodoin, "MIO-TCD: A new benchmark dataset for vehicle classification and localization," *IEEE Transactions on Image Processing*, 2018.

[19] K. Abdulrahim and R. A. Salam, "A new motion segmentation technique using foreground-background bimodal," *Malaysian Journal of Science Health & Technology*, vol. 2, no. 1, 2018.

Obaida M. Al-Hazaimeh is an associate professor of network security and image processing in Computer Science and Information Technology Department, Al-Huson University College, Al-Balqa' Applied University. He holds a PhD in computer science. He co-authored 33 research articles in leading ISI / international refereed journals.

Ma'moun Al-Smadi is a lecturer in Electrical and Electronics Engineering Department, Al-Huson University College, Al-Balqa' Applied University. He received the BS degree in electrical engineering/computer and MSc degree in electrical engineering/control and power both from Jordan University of Science and Technology, Irbid, Jordan in 1998, and 2003, respectively. His research interests are in the areas of digital image processing, computer vision, and machine learning. He co-authored 6 articles in leading International refereed journals.