An Investigation of a Convolution Neural Network Architecture for Detecting Distracted Pedestrians

Igor Grishchenko[1], El Sayed Mahmoud[2]
Faculty of Applied Science and Technology
Sheridan College, Oakville, Canada

***Abstract*—The risk of pedestrian accidents has increased due to the distracted walking increase. The research in the autonomous vehicles industry aims to minimize this risk by enhancing the route planning to produce safer routes. Detecting distracted pedestrians plays a significant role in identifying safer routes and hence decreases pedestrian accident risk. Thus, this research aims to investigate how to use the convolutional neural networks for building an algorithm that significantly improves the accuracy of detecting distracted pedestrians based on gathered cues. Particularly, this research involves the analysis of pedestrian' images to identify distracted pedestrians who are not paying attention when crossing the road. This work tested three different architectures of convolutional neural networks. These architectures are Basic, Deep, and AlexNet. The performance of the three architectures was evaluated based on two datasets. The first is a new training dataset called SCIT and created by this work based on recorded videos of volunteers from Sheridan College Institute of Technology. The second is a public dataset called PETA, which was made up of images with various resolutions. The ConvNet model with the Deep architecture outperformed the Basic and AlexNet architectures in detecting distracted pedestrian.**

Keywords*—Convolutional neural networks; computer vision; cognitive load; distractive behavior

I. Introduction

Pedestrians are the most vulnerable objects observed by autonomous vehicles because they travel along streets, roads, sidewalks, alone and with others in both busy and idle areas. According to the Canadian Motor Vehicle Traffic Collision Statistics 2015 [1], pedestrians accounted for 15.4% of fatalities and 14.3% of serious injuries in all motor vehicle accidents. Pedestrians can make changes in their path because many roads and streets cannot have physical constraints that ensure pedestrians use the appropriate behavior all the time. This makes planning a safe route challenging even with all the current technologies equipped to self-driving cars today.

One of the main reasons for the difficulties in detecting and predicting pedestrian behavior is attributed to the use of mobile devices while walking. Pedestrians who use handheld devices tend to walk blindly into the path of a moving vehicle. Doing so increases the likelihood of a collision. Using devices while walking limits pedestrian cognitive functions which in turn could lead to walking with high risk to cause the accident [2].

The use of handheld devices by pedestrians affects their cognitive load and the ability to pay close attention to the road, thus, increases the car accident risk. This creates a further challenge for the self-driving car to plan the safest route because the walking path of a distracted pedestrian is not related to the current road conditions. Identifying pedestrians who use cell phones during their walking will significantly decrease the number of injuries and deaths due to distracted pedestrians. This study developed and trained a Convolutional Neural Network (CNN) to detect pedestrians who use handheld devices while crossing the road. Ultimately, this work developed the distracted pedestrian detector, based on convolutional neural networks, which is able to analyze whether the pedestrian is distracted or not in real-time.

A. Motivation

The motivation of this thesis is to improve the safety of pedestrians by leveraging the convolutional neural networks. The application of convolutional neural networks could improve the accuracy of detecting pedestrians and identify if they are distracted. The ConvNet investigates image structural information and builds the neural network model in a more insightful manner than non-deep neural networks [3]. Today, research on the detection of pedestrian motions and route planning is conducted frequently with many readily available publications. However, only few mention the fact that pedestrians can be distracted and how their behavior and movement can and may change unexpectedly due to cognitive dissonance. This study investigates the problem of distracted pedestrians by implementing the detector based on the ConvNets, which can identify whether the observed pedestrian is holding the handheld device or not. Stakeholders who benefit from the proposed algorithm are the vehicle manufactures, smart cities project teams, and researchers. As mentioned previously, drivers are also distracted by handheld devices as well. Thus, the developed algorithm can also be applied to warn a driver if a pedestrian is distracted and the chance of accident will overall decrease. The main goal of this work is to improve the accuracy of automated vehicles to make their choices safer and minimize the possibility of injury.

B. Organization of Paper

The rest of this paper is organized as follows, the literature review chapter covers prior researches related to autonomous vehicles and the detection of pedestrians by examining different techniques such as neural nets (MLP), knowledge extractions, and model tuning. It consists of studies that focus on human cognition research and how handheld devices can lead to unwilling motions while walking. The methodology chapter focuses on describing what methodology was used and how it was applied in detail. This involves the selection of

ConvNet architecture, model training and tuning as well as testing the detector on the videos of participants. Lastly, the results chapter presents the gathered experimental findings, a review of the findings with analysis and future research opportunities.

II. Literature Review

With the sharp growth of self-driving cars in the automotive industry and the increasing usage of handheld devices by pedestrians, the ability for autonomous vehicles to detect distracted pedestrians has become prevalent, hence receiving a considerable amount of attention and extensive research on determining whether the pedestrian is distracted or not [3] [4].

Many research groups concentrated on the challenge of determining the limb positioning of a pedestrian for a long time and introduced a variety of models. Some studies applied classical machine learning algorithms by fitting labeled data into models, such as Gaussian process (GP) regression [5], Support Vector Machines (SVM) [6], and Mixed Markov-Chain Model (MMCM) [7]. Other groups conducted research considering deep neural networks. Dominguez-Sanchez et al. conducted research for the improvement of pedestrians' motions detection by leveraging convolutional neural network (CNN) [3]. Another approach proposed by Yamashita et al. involves the use of Multi-Task Convolutional Neural Network for the detection of pedestrians and the position of their limbs simultaneously [8]. The latter two approaches will be considered the closest to this study and will be the focus of this study's research.

It is essential to detect distracted pedestrians since it can help to prevent vehicle conflicts and reduce vehicle traffic due to indecisions when crossing and overall slower crossing speed [9]. According to Zaki et al., this type of research would benefit multiple domains which include road safety which extends the application of computer vision (CV). The potential improvement of the current methodology for identifying distracted pedestrians would be the exploration of head and hands positional tracking [9].

With the growth of autonomous cars in the motor vehicle industry and the increasing number of distracted pedestrians, the importance of this research as well as the understanding and analysis of the distracted walking behavior of pedestrians have been more than reaffirmed. Recent studies about the exploration of pedestrians' gait benchmarks for the identification of whether they are distracted or not has been completed [9].

A survey of theory and practice in the interaction between self-driving cars and pedestrians conducted by Rasouli et al. showed that pedestrians who are distracted by handheld devices are 75% more likely to display unintentional blindness [10]. Another study conducted by Neider et al. investigated that distraction arising from the cell phone usage challenges pedestrians' ability to estimate the time-to-contact of traffic accurately, which increases the odds of failing to cross a road safely. Fig. 1 visualizes the results gathered by Neider et al. during the research experiments and shows the percentage of attempts in which participants successfully crossed the street [11]. Fig. 1 demonstrates that pedestrians who were talking on the phone while crossing the street were less likely to successfully cross the road compared to non-distracted pedestrians [11].

Distracted pedestrians tend to change their walking direction more often and on average, cross the street slower than undistracted pedestrians, which can lead to unwilling accidents [10] [9]. The ability of autonomous cars to detect pedestrians who are not paying attention while crossing the road can improve road safety. Since the motor vehicle industry is steadily shifting towards self-driving cars, these autonomous cars must recognize if a pedestrian is not paying attention to the road, in order to prevent any hazards associated with distraction [12]. Current studies focus on analyzing pose and extracting gait parameters of pedestrians to determine whether the pedestrian(s) is distracted or not [12] [9].

This study's intention is to improve self-driving cars' accuracy in collisions detection and path planning by identifying whether the pedestrians are distracted or not. The main goal of this work is to use a convolutional neural network model to detect distracted pedestrians by examining specific distracted behavior scenarios of pedestrians.

A. Convolutional Neural Networks in Computer Vision

Deep Convolutional Neural Networks (ConvNet) has demonstrated amazing performance in several computer vision tasks, including face recognition, digits recognition, and image classification, due to the ability to extract visual benchmarks from the pixel-level content [13]. However, it was a great challenge to train the deep ConvNets due to the lack of training data and computational power in the past, but many methods had been proposed to overcome this problem since 2006 [14]. In 2012, Krizhevsky et al. proposed a classic ConvNet architecture, AlexNet, and demonstrated notable improvements in the image classification tasks [15]. AlexNet showed high levels of accuracy in image recognition applications and received considerable attention from the community, and therefore, many studies were conducted to improve or even surpass AlexNet's performance. Subsequently, more effective and deeper ConvNet architectures were proposed: ZFNet, VGGNet, GoogleNet, and ResNet [14]. The typical modification of these new architectures was the increased depth in order to extract even more features from the input. Furthermore, deep ConvNets were successfully applied for pedestrians' detection problems by estimating the movement of their limbs [16] [3].

The research by Lu et al. examined the application of convolutional neural networks for player detection and team classification in group sports such as basketball, ice hockey, and soccer from broadcasting videos [17]. They also experimented on a pedestrian dataset to evaluate the generality of their approach. Their model performed very well and was able to classify each team in different sports with 97% accuracy. Table I shows the confusion matrix of the percentage of players being classified by teams in the 4 different data sets [17]. Table I represents the proportion of players in each team being classified into the corresponding team. Classes TA, TB, and O refer to Team A, Team B, and Others accordingly.

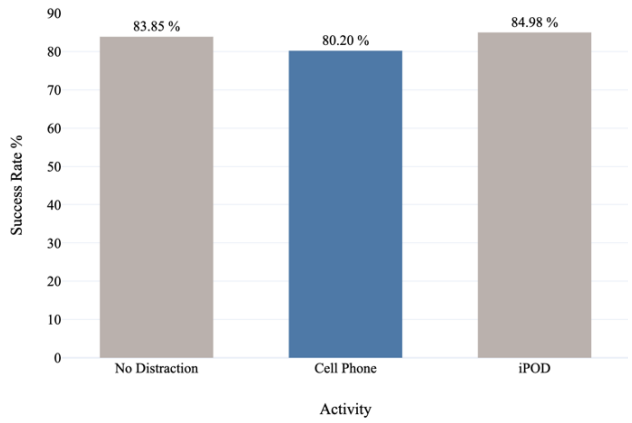

Fig. 1. The Percentage of Pedestrians' Success Crossing While being Distracted or not.

The study conducted by Dominguez-Sanchez et al. evaluated the ability and performance of the current convolutional neural networks and proved how CNNs can impressively perform an estimation task of determining limbs movement of a pedestrian. During the research, they trained their networks with their own novel video dataset which was processed into frames through the image preprocessing pipeline. Only one of every six frames were used for the input during experiments of pedestrians' limb position and movement detection. After the evaluation of AlexNet, GoogleNet, and ResNet architectures, they identified that ResNet was the best for pedestrians' movement recognition and demonstrated 79% accuracy in the test set. Table II illustrates the results obtained by the ResNet in the test set [3].

TABLE. I. CONFUSION MATRIX OF TEAM MEMBERSHIP CLASSIFICATION IN 4 DATASETS

Dataset	Classes	**TA**	**TB**	**O**
Basketball	TA	**99.65**	0.18	0.17
	TB	0.91	**97.88**	1.21
	O	0.86	1.71	**97.43**
Ice Hockey	TA	**98.91**	0.72	0.37
	TB	1.33	**97.99**	0.68
	O	0.69	1.36	**97.95**
Soccer Set 1	TA	**98.63**	0.24	1.13
	TB	0.83	**98.23**	0.94
	O	2.08	1.41	**96.51**
Soccer Set 2	TA	**98.33**	0.44	1.23
	TB	0.91	**97.78**	1.31
	O	2.46	1.37	**96.17**

TABLE. II. CONFUSION MATRIX OF RESULTS OBTAINED BY RESNET

	Front	**Left**	**Right**
Front	**0.980**	0.011	0.008
Left	0.058	**0.841**	0.100
Right	0.081	0.265	**0.652**

Abdulnabi et al. introduced a modified deep convolutional neural network architecture that enables multitasking, so different CNNs can share knowledge among each other [18]. Their learned Multi-Task CNN demonstrated better performance in predicting semantic binary attributes by sharing visual knowledge between tasks. The results obtained from experiments on two different datasets and multiple different CNNs shows that Multi-Task CNN used by Abdulnabi et al. outperformed single-task neural networks and achieved 92% accuracy in attribute predictions in images [18] Deep convolutional neural networks demonstrated amazing performance in pedestrians and attribute detection and were selected as the approach for this research.

III. DATA SOURCES

One of the data sources used in this research was built by recording student volunteers from the Sheridan College Institute of Technology (SCIT dataset). Recording students' videos to create a dataset was approved by the Sheridan Research Ethics Board. The total number of participants was 15 with different demographics such as gender, race, and age which allowed us to construct a good quality diverse dataset. The videos were recorded in an enclosed environment where each participant was asked to mimic a distracted/non-distracted pedestrian, based on the attributes listed in Table III while crossing the road. These video recordings of their walk were incorporated into the training set and further used for this study. The volunteers were recorded from three different positions for both front and rear views in order to capture every possible angle, direction, and position. Then, all the video footage was split into frames and labeled based on the participants' behavior to differentiate distracted and non-distracted scenarios. Each participant had around 350 frames per each activity, thus, we formed $350 \times 15 \approx 5{,}000$ images per activity after data preprocessing.

Another data source was built from Composition of PEdesTrian Attribute (PETA) dataset with 19000 images, with the image size ranging from 17×39 pixels to 169×365 pixels, which were released by Deng et al. during their research [19]. They also provided attribute annotations for each image in order to perform benchmarks detection. Yet, their dataset did not provide any labels whether the person on an image is distracted or not. Thus, all the images were reviewed and classified manually to fit the purpose of this research.

A. Determining Pedestrians Distracted behavior Scenarios

After collecting data based on walking pedestrians, all the images were broken down into two classes: *distracted* and *non-distracted* pedestrians. The literature has been explored to identify what type of behavior can cause cognitive load and result in an unsafe road crossing. According to research conducted by Mwakalonge et al., 75% of pedestrians who were walking while taking on a cell phone displayed inattention blindness and failed to notice unusual activity [20]. Another study by Neider et al. performed the experiment in a virtual pedestrian environment and determined that participants who were distracted by music or texting were more likely to be hit by an automobile [5]. 5 different scenarios were identified where a pedestrian is considered to be distracted based on their hands and head positioning. Table III provides an overview of

those scenarios as well as example images from the SCIT dataset. Then, PETA dataset images that fall under the identified scenarios were manually moved to a different directory to be separated from the images that were identified as non-distracted pedestrians. As for the SCIT dataset, all the videotaped volunteers were asked to mimic distracted and non-distracted behavior before the recording, thus, all the data were already structured and easily distributed in two classes. Also, each distracted and non-distracted scenario was recorded from different views to simulate real-life situations as much as possible.

IV. Methodology

The development phases for the proposed detector include: (i) identifying the appropriate sample size to train an accurate ConvNet image recognition classifier, (ii) datasets preprocessing to improve the quality of the data, and (iii) designing a ConvNet architecture and fine-tuning hyper-parameters to get the accurate classifier.

A. Identifying Appropriate Sample Size

The most effective dataset size to accurately train a ConvNet model is determined iteratively and can be guided by the distribution of classes and their behaviors. Therefore, it is not clearly defined which sample size would to train an accurate ConvNet pedestrian classifier. Li et al. used the Caltech-101 dataset which contains 9,144 images with a variety of classes to train and test their CNN image classifier and achieved 89% accuracy [21]. The samples of 4,000 images and 30,000 images of distracted and non-distracted pedestrians were gathered from the PETA and SCIT datasets accordingly. However, the whole number of images in the SCIT dataset was not used in the experiments since this number is calculated based on the number of images for each behavior example where we have 5,000 images per scenario. Therefore, we used all the images from the non-distracted scenario set to create the first-class and randomly selected 1,000 images from each of the distracted scenarios sets to create the second class. Eventually, we constructed the dataset of 10,000 images of distracted and non-distracted classes based on the SCIT data.

B. Preprocessing of SCIT and PETA datasets

Before training the detector and conducting different experiments, people were cropped from the frames in the SCIT dataset gathered by our experiment. A pretrained Mask R-CNN object detector was used to detect people in each image and annotate their bounding boxes to perform the cropping. The resolution of the cropped pedestrian images is ranging from 62 × 224 pixels to 494 × 987 pixels in the SCIT dataset. The amount of blur in each image was also computed in order to remove images with excessive amounts of blurring that improved the dataset quality. Further, data augmentation techniques were applied to both PETA and SCIT datasets in order to increase the size of the datasets. Particularly, we augmented our data by rescaling, zoom-range, and fill-mode.

C. Determining CNN Architecture and Fine-Tuning

Convolutional Neural Networks have been selected due to their convolution layers which extract features from an input image and learn from them by exploiting small chunks of input data in order to preserve the spatial relationship between them.

TABLE. III. Description of Scenarios when Walking Pedestrian is Distracted

Scenario Description
Head down and holding the phone with the left hand. A participant is chatting on the phone.
Head down and holding the phone with the right hand. A participant is chatting on the phone.
Head down and holding the phone with both hands. A participant is chatting on the phone.
The left hand is near the head. A participant is speaking over the phone.
The right hand is near the head. A participant is speaking over the phone.

We proposed two architectures Basic and Deep with 3 and 5 convolutional layers accordingly to undertake the problem of distracted pedestrian detection:

The first architecture has the following structure: The first convolutional layer has 16 filters of size 3 with ReLU activation function followed by batch normalization and max-pooling layer of size 2×2; the second convolutional layer has 32 filters of size 3 with Tanh activation function followed by batch normalization and max-pooling layer of size 2×2; the third convolutional layer has 64 filters of size 3 with ReLU activation function followed by batch normalization and max-pooling layer of size 2×2. The last max-pooling layer is followed by the dropout layer with a 25% dropout rate. After the aforementioned layers, we have flatten layer followed by two dense which also called fully connected layers. The first dense layer has 64 nodes with the ReLU activation function and the second has only 2 nodes with Sigmoid activation function since we need to find a probability of the pedestrian being distracted or not. This architecture is presented on the left side of Fig. 2.

The second architecture is the modification of the above one where the second and third layers were duplicated such that two convolutional layers are stacked together before every max-pooling layer. Multiple stacked convolutional layers can be able to learn more complex features from the input before the destructive max-pooling layer [22]. We considered this technique to be promising in the detection of distracted pedestrian problem. The second architecture is shown on the right side of Fig. 2.

We applied the same hyper-parameters to both architectures; we used RMSprop optimizer with default parameters: learning rate = 0.001 and $\beta = 0.9$. The loss function we selected was the binary cross-entropy since this function better suits classification tasks with 2 classes [23]. All the convolutional layers were preceded by the zero or "same" padding to preserve the size of post convolution. Finally, we applied the early stopping regularization technique to prevent the model from overfitting.

D. Testing Strategy

The detector was tested with randomly selected images of distracted and non-distracted pedestrians which have not been seen by the model during training. Since SCIT data consists of 15 different participants, we randomly selected 4 participants and their images to generate the test set. The data of the other 11 participants were used for training. This 11/4 split is

equivalent to a 75/25 data split, where 75% of data was used to train the model and the other 25% was used to test the model. This approach allowed us to always test our model on the people's data which the model had never seen before. Regarding the PETA dataset, since most of its data points represent a unique pedestrian, we randomly split data following the same 75/25 approach. Besides, the data in both datasets was always shuffled every time when we trained a new version of the model in order to reduce variance, make sure that the model remains general, and prevent overfitting. We conducted an experiment to examine how both our architectures can perform on different combinations of datasets, which drastically different in the resolution of the images. AlexNet architecture was also evaluated on the same datasets to compare it with our proposed architectures.

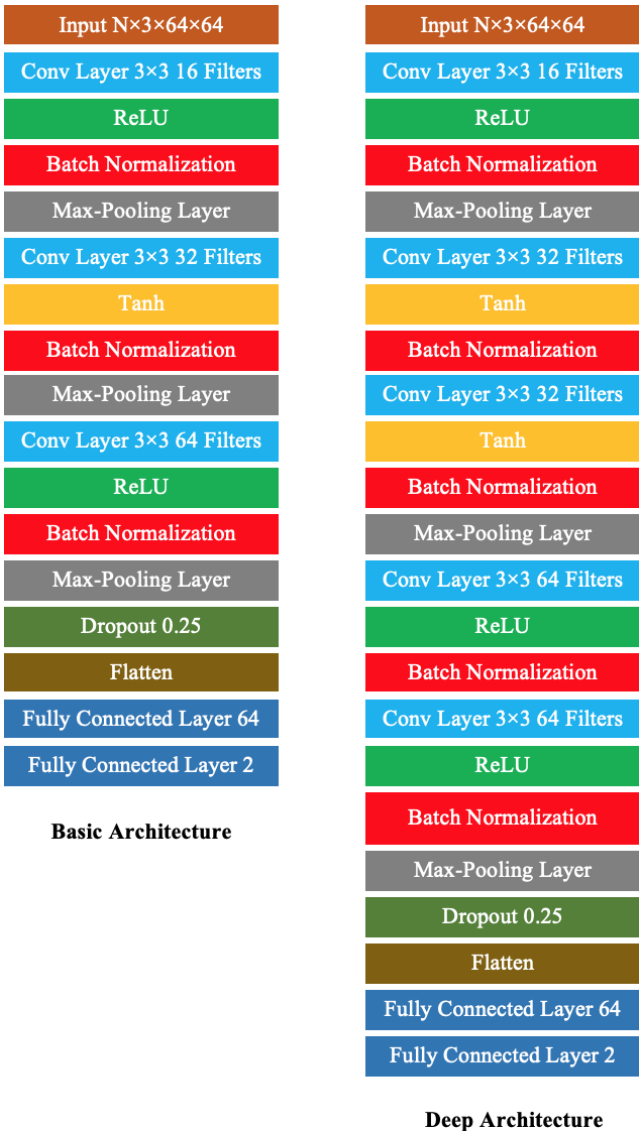

Fig. 2. Architectures of Distracted Pedestrians Detector.

E. Proposed Experiment

The purpose of the experiment was to see how the quality of the images would affect the performance of the ConvNet based on different architectures. Therefore, we created three different sample sets from the SCIT and PETA datasets for this test. The first sample was made of only the SCIT dataset where all the images had high resolution (62 × 224 pixels to 494 × 987 pixels) and distraction scenarios were equally distributed. The second sample was constructed from the PETA dataset and its images had a relatively low number of pixels (17 × 39 pixels to 169 × 365 pixels). The third data sample was created using both SCIT and PETA dataset where high and low image resolution (17 × 39 pixels to 494 × 987 pixels) were combined. The purpose of the third sample was to see whether the ConvNet accuracy would degrade or not if we feed data to it which has a huge range in quality to it.

The models with Basic and Deep architectures were trained and tested on the aforementioned datasets. We also investigated how AlexNet architecture that achieved state-of-the-art results in many computer vision tasks would tackle the distracted pedestrian detection problem [24]. AlexNet is a much deeper network with more filters in each convolutional layer. The model with AlexNet architecture was also trained on the same data samples, so we could compare its performance with our Basic and Deep architectures. The reason why the AlexNet had been also evaluated was to examine if the deeper network with more filters would be smarter in the feature extraction related to our problem and would have better accuracy in distracted pedestrian detection. Fig. 3 illustrates the design of the experiment.

V. Results and Analysis

This section shows the experimental results of building the Distracted Pedestrian Detector based on different combinations of the datasets: SCIT, PETA, and a combination of both. This work tested two different ConvNet architectures. The first is called Basic, and the second is called Deep, which duplicates the second and third layers of the Basic architecture. Additionally, we examined how the AlexNet model would tackle the distracted pedestrian detection problem based on the combinations of the aforementioned datasets. The Deep ConvNet architecture was more efficient than the Basic and AlexNet architectures in detecting the distracted pedestrians based on all three datasets.

A. Effect of the Image Resolution on the Performance

The highest accuracy of the Distracted Pedestrian Detector with Deep architecture for the SCIT dataset was 95.11%. Fig. 4 shows the average accuracies of the Deep, Basic, and AlexNet architectures trained and tested on the SCIT data sample. Since the SCIT datasets had the highest resolution, this particular evaluation demonstrates how the architectures behave on images with a big number of pixels. The Deep architecture also showed the highest average 94.02% accuracy. The Basic architecture was the second in the accuracy and achieved 90.00% on average. Lastly, the performance of AlexNet was close to the Basic architecture but demonstrated lower average accuracy – 89.23%. Based on the high precision and recall scores, shown in Table IV, we can see that all the models trained on the SCIT data were able to correctly classify a high number of the relative data points. This is supported by the f1 score since it was also relatively high too, meaning that models were general and unbiased. This was due to the SCIT dataset being well distributed and provided the models with balanced

training and testing data. We can see that all the architectures performed relatively well on the dataset which contains images with high resolution.

When trained and validated on the PETA dataset, all the architectures demonstrated lower accuracies. This can be explained by a certainly low resolution of images in the PETA dataset. Fig. 5 visualizes the average accuracies achieved by the Deep, Basic, and AlexNet architectures trained and tested on the PETA data sample. The Deep architecture maintained the first place and showed an average 85.44% accuracy. The AlexNet architecture had the 83.67% accuracy on average what was close to the Deep one. Yet, the Basic architecture demonstrated the biggest reduction in accuracy and achieved 78.01% what notably different from the score of Deep and AlexNet architectures. The Basic ConvNet had the smallest number of convolutional layers and, therefore, the minimum number of filters. It performed relatively bad in distinguishing between distracted and non-distracted pedestrians. We tried to increase the number of filters in each convolutional layer by 4 times such that it had 64 filters in the first layer, 128 filters in the second layer, and 256 in the third layer. Unfortunately, this only worsened the architecture, because the high number of filters caused model overfitting since the training accuracy was 97.15% while the validation accuracy was only 76.34%. This indicates that the three convolutional layers are not enough to deal with images with a small number of pixels.

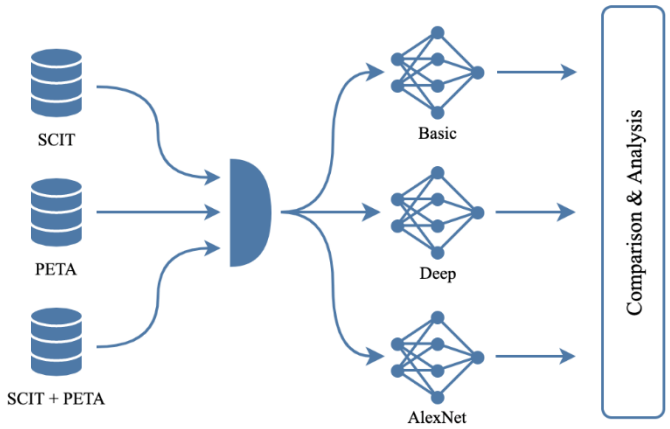

Fig. 3. Experiment Design.

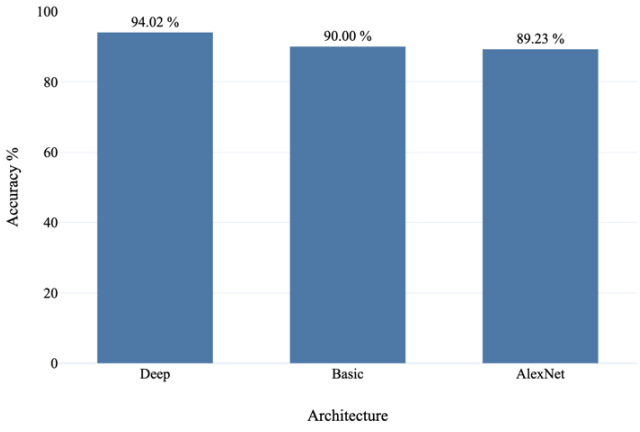

Fig. 4. Average Accuracy of Architectures for SCIT Dataset.

TABLE. IV. AVERAGE PRECISION, RECALL, AND F1 SCORE METRICS OF MODELS FOR SCIT DATASET

	Precision	Recall	F1 Score
Deep	0.9434	0.9374	0.9403
Basic	0.9246	0.8812	0.9024
AlexNet	0.8826	0.9000	0.8912

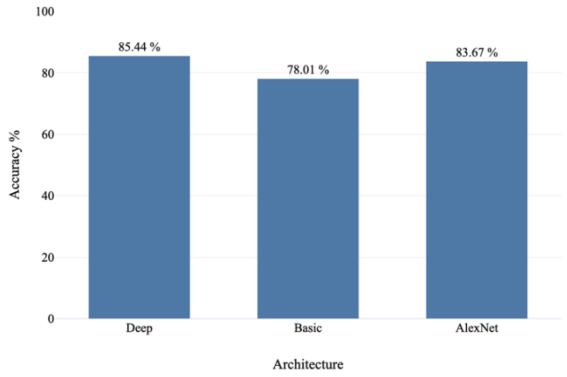

Fig. 5. Average Accuracy of Architectures for PETA Dataset.

If we analyze the precision, recall, and f score metrics, demonstrated in Table V, we can see that the recall metric significantly dropped compared to the precision metric. It means that the models evaluated on the PETA data classified more distracted pedestrians as non-distracted. We then can conclude that the data with low-quality images did not allow models to learn enough patterns, since it was relative to the distracted behavior. Also, some of the images were captured from a distance making it really difficult for the models to detect if an observed pedestrian is holding a handheld device or not.

The third dataset which was used for the evaluation of the architectures was the combination of both SCIT and PETA data. The highest accuracy was demonstrated by the Deep architecture which achieved 88.78%. The average accuracy of the Deep, Basic, and AlexNet architectures trained and evaluated on the combination of SCIT and PETA datasets is shown in Fig. 6. The Deep architecture, again, showed the best average accuracy – 87.01%. The accuracies of AlexNet and Basic architectures were 84.32% and 80.56%, respectively. All the architectures did not improve much, and their average accuracies were approximately 2% better compared with the models trained and tested only on the PETA dataset. These results illustrate that even if we combine the images with low and high resolutions, the images with a low number of pixels in the set still affects the ability of ConvNet accurately detect distracted pedestrians. Besides, the big range of the resolution could also be a reason for the not significant improvement of the architectures. ConvNets could not establish a clear pattern from the extracted features to find the difference between distracted and non-distracted scenarios.

Table VI shows the precision, recall, and f1 score metrics obtained by the ConvNet models trained and tested on the combination of both SCIT and PETA datasets. It is clear that if we add high-quality images to the dataset that contains images with a low number of pixels, the models can learn more features and distinguish distracted and non-distracted

pedestrians with better accuracy. However, the following metrics are still lower compared to the obtained metrics in Table IV, which demonstrates again, that data with low-quality images has a big influence on the architectures, even if data points with a big number of pixels are dominant in this dataset.

Since the model based on the Deep architecture demonstrated higher accuracy across all three datasets, the one-way analysis of variance (ANOVA) test was used to determine if the Deep architecture's score is significantly different from the Basic and AlexNet models. The ANOVA test was conducted on three different sets of models trained on the different datasets as shown in Fig. 4, Fig. 5, and Fig. 6. The *p-value* from the three test results was the following: 0.00003, 0.000025, 0.000027 for the sets of models trained on the SCIT, PETA, and combination of SCIT and PETA datasets, accordingly. Since the *p-value* across all the datasets was less than 0.05, this indicates that the models' accuracies were significantly different and not from the same. Thus, we can conclude that the difference in the model's scores is significant showing that the Deep actually had the highest accuracy.

B. Impact of Architecture Design

We also inspected the filters and feature maps during the layers' convolution of Basic and Deep ConvNet architectures. Since Deep architecture was designed to have the second and third layers combined together followed by the max-pooling layer, the third layer was able to receive a more precise feature map where we still can recognize the original image as shown in Fig. 7. In contrast, all the convolutional layers in Basic architecture are split by max-pooling layer, therefore, the feature map of the third layer in the Basic architecture is less interpretable and contains high-level concepts as displayed in Fig. 8. From Fig. 7 and Fig. 8, we can see that the feature map in the third convolutional layer of the Deep architecture still contains visual concepts like edges, which are useful for our problem since the detector needs to evaluate the position of the pedestrian limbs to differentiate distracted and non-distracted behavior. While the feature map in the third layer of the Basic architecture looks more like the abstraction of the original image and contains high-level features that might have more information about small parts of the image such as a mobile device in the hands. Of course, both low and high-level features are highly important to accurately detect distracted pedestrians. Though, the design of the Deep architecture allowed filters to extract more low-level features that helped ConvNet to characterize the position of pedestrian limbs and better recognize the distractive action. This explains why ConvNet with Deep architecture outperformed the Basic ConvNet across all the three datasets since the Basic architecture could not extract enough features related to the pedestrians' actions.

TABLE. V. AVERAGE PRECISION, RECALL, AND F1 SCORE METRICS OF MODELS FOR PETA DATASET

	Precision	Recall	F1 Score
Deep	0.8990	0.8251	0.8604
Basic	0.8780	0.7341	0.7996
AlexNet	0.8915	0.8024	0.8446

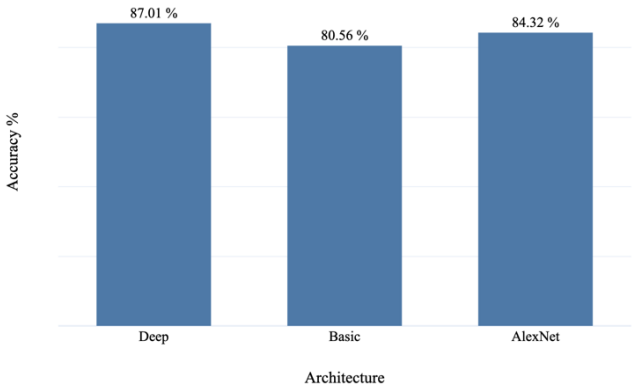

Fig. 6. Average Accuracy of Models for SCIT and PETA Datasets.

TABLE. VI. AVERAGE PRECISION, RECALL, AND F1 SCORE METRICS FOR THE COMBINATION OF SCIT AND PETA DATASETS

	Precision	Recall	F1 Score
Deep	0.8888	0.8566	0.8724
Basic	0.8272	0.7929	0.8097
AlexNet	0.8685	0.8266	0.8470

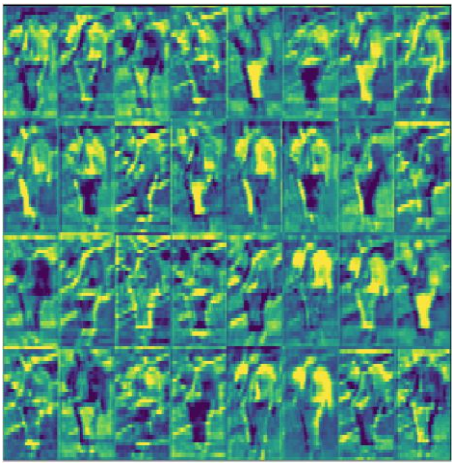

Fig. 7. Visualization of the Filters in the Third Conv Layer of the Deep Architecture.

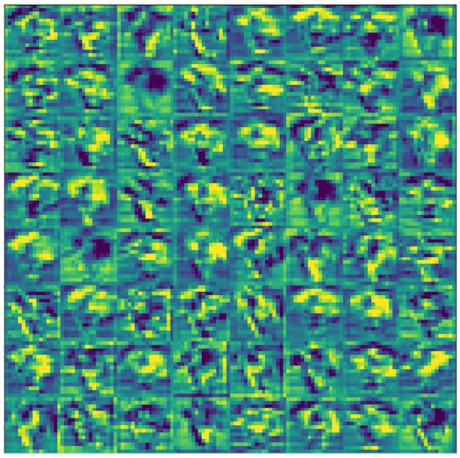

Fig. 8. Visualization of the Filters in the Third Conv Layer of the basic Architecture.

Interestingly enough that Deep and AlexNet architectures had a similar design in terms of the combined convolutional layers. While the Deep architecture combined the second with third and the fourth with fifth layers, the AlexNet architecture design combined the third, fourth, and fifth convolutional layers without max-pooling layers between them. But based on the gathered results demonstrated above, the Deep architecture achieved higher average accuracies across all the three datasets. Despite the fact, that even if AlexNet has a similar structure to the Deep architecture, its combined convolutional layers focused mostly on the extraction of the high-level features since they were the last group and received feature maps that already got through multiple max-pooling layers. Therefore, AlexNet could not extract more low-level features like the Deep architecture. This derives the conclusion that the low-level features which are responsible for the detection of edges and shapes played a very important role in the distracted pedestrian detection problem and allowed the Deep architecture to outperform the AlexNet and Basic ConvNets.

VI. Conclusion

This research aimed to explore the application of convolutional neural networks to address the problem of detecting distracted pedestrians automatically. This work investigated various combinations of CNN architectures and datasets to build an effective distracted pedestrian detector. A novel training dataset was created from video recordings of volunteer participants from the Sheridan College Institute of Technology when they acted as distracted and non-distracted pedestrians. This dataset is called SCIT and could be used for further research in various computer vision research problems related to human detection. Three ConvNet models were implemented with different architectures: Basic, Deep, and AlexNet. Each model was trained and tested on three different datasets: SCIT, PETA, and the combination of both. The results from the experiment had indicated that the model that utilized the Deep architecture had outperformed the other models that used the Basic and AlexNet architectures when applied to all the datasets. The developed detector could be used for autonomous vehicles and driver alert systems to identify distracted pedestrians who cross the street and minimize the probability of injury. The detector would also be useful for the variety of stakeholders including the vehicle manufactures, researchers, and smart cities project teams.

VII. Future Work

The detector currently takes an entire image and makes a prediction based on the extracted features. The next step will be to modify the algorithm so that it would extract pedestrian limbs such as head and hands from each image and evaluate them independently instead of analyzing a complete image. This modification will increase the efficiency of the system because it will minimize the misclassification of handheld devices with other potential objects in the pedestrian's hands. An analysis of how a pedestrian's head direction changes would also create a meaningful impact on when identifying if a pedestrian is distracted.

Predicting the route of a distracted pedestrian will be another perspective addition to the system. Distracted pedestrians tend to change their route unexpectedly what increases the possibility of an accident. With the knowledge that a pedestrian is distracted, his/her long-term path could be predicted more accurately. The information about pedestrians' future path and if they are distracted or not could advance the safe route planning for self-driving cars.

Sequential frame classification can be another improvement to the detector. In this case, extraction of the sequence features, which are also called temporal or time-related features, will be required in addition to the features of the images. This approach could help identify when a pedestrian had acted similar to a distracting behavior for a short period of time when the pedestrian's action was not an actual distraction. This could reduce the number of false positives that would improve the reliability of the detector.

Acknowledgment

This work has resulted in Igor Grishchenko's undergraduate thesis of the Honours Bachelor of Computer Science (Mobile Computing). The collection of data from human participants was approved by the Sheridan Research Ethics Board. We are grateful for the insightful feedback on the work received from the thesis advisory committee members at Sheridan College.

References

[1] Transport Canada, "Canadian Motor Vehicle Traffic Collision Statistics: 2017," Transport Canada, 27-Feb-2019. [Online]. Available: https://www.tc.gc.ca/eng/motorvehiclesafety/canadian-motor-vehicle-traffic-collision-statistics-2017.html. [Accessed: 18-Nov-2019].

[2] G. Yogev-Seligmann, J. M. Hausdorff, and N. Giladi, "Do we always prioritize balance when walking? Towards an integrated model of task prioritization," Movement Disorders, vol. 27, no. 6, pp. 765–770, 2012.

[3] A. Dominguez-Sanchez, M. Cazorla, and S. Orts-Escolano, "Pedestrian Movement Direction Recognition Using Convolutional Neural Networks," IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 12, pp. 3540–3548, 2017.

[4] Y. Tang, L. Ma, W. Liu, and W.-S. Zheng, "Long-Term Human Motion Prediction by Modeling Motion Context and Enhancing Motion Dynamics," Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018.

[5] Y. Chen, M. Liu, S.-Y. Liu, J. Miller, and J. P. How, "Predictive Modeling of Pedestrian Motion Patterns with Bayesian Nonparametrics," AIAA Guidance, Navigation, and Control Conference, 2016.

[6] J.-T. Wang, D.-B. Chen, H.-Y. Chen, and J.-Y. Yang, "On pedestrian detection and tracking in infrared videos," Pattern Recognition Letters, vol. 33, no. 6, pp. 775–785, 2012.

[7] A. Asahara, K. Maruyama, A. Sato, and K. Seto, "Pedestrian-movement prediction based on mixed Markov-chain model," Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS 11, 2011.

[8] T. Yamashita, H. Fukui, Y. Yamauchi, and H. Fujiyoshi, "Pedestrian and part position detection using a regression-based multiple task deep convolutional neural network," 2016 23rd International Conference on Pattern Recognition (ICPR), 2016.

[9] M. H. Zaki and T. Sayed, "Exploring walking gait features for the automated recognition of distracted pedestrians," IET Intelligent Transport Systems, vol. 10, no. 2, pp. 106–113, Jan. 2016.

[10] A. Rasouli and J. K. Tsotsos, "Autonomous Vehicles That Interact With Pedestrians: A Survey of Theory and Practice," IEEE Transactions on Intelligent Transportation Systems, pp. 1–19, 2019.

[11] M. B. Neider, J. S. Mccarley, J. A. Crowell, H. Kaczmarski, and A. F. Kramer, "Pedestrians, vehicles, and cell phones," Accident Analysis & Prevention, vol. 42, no. 2, pp. 589–594, 2010.

[12] A. Rangesh, E. Ohn-Bar, K. Yuen, and M. M. Trivedi, "Pedestrians and their phones - detecting phone-based activities of pedestrians for

autonomous vehicles," 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), 2016.

[13] D. Tomè, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, and S. Tubaro, "Deep Convolutional Neural Networks for pedestrian detection," Signal Processing: Image Communication, vol. 47, pp. 482–489, 2016.

[14] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," Pattern Recognition, vol. 77, pp. 354–377, 2018.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, Nov. 2015.

[16] Y.-L. Hou, Y. Song, X. Hao, Y. Shen, and M. Qian, "Multispectral pedestrian detection based on deep convolutional neural networks," 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), 2017.

[17] K. Lu, J. Chen, J. J. Little, and H. Hea, "Lightweight convolutional neural networks for player detection and classification," Computer Vision and Image Understanding, vol. 172, pp. 77–87, 2018.

[18] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-Task CNN Model for Attribute Prediction," IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 1949–1959, 2015.

[19] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian Attribute Recognition At Far Distance," Proceedings of the ACM International Conference on Multimedia - MM 14, 2014.

[20] J. Mwakalonge, S. Siuhi, and J. White, "Distracted walking: Examining the extent to pedestrian safety problems," Journal of Traffic and Transportation Engineering (English Edition), vol. 2, no. 5, pp. 327–337, 2015.

[21] Q. Li, Q. Peng, and C. Yan, "Multiple VLAD Encoding of CNNs for Image Classification," Computing in Science & Engineering, vol. 20, no. 2, pp. 52–63, 2018.

[22] J. B. Ahire, Artificial Neural Networks: The brain behind AI. 2018.

[23] P. Lakhani, D. L. Gray, C. R. Pett, P. Nagy, and G. Shih, "Hello World Deep Learning in Medical Imaging," Journal of Digital Imaging, vol. 31, no. 3, pp. 283–289, Mar. 2018.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.