Human Crawl vs Animal Movement and Person with Object Classifications Using CNN for Side-view Images from Camera

Tarun Choubisa, Mohan Kashyap, Kodur Krishna Chaitanya
Dept. of Electrical Communication Engg., Indian Institute of Science, Bengaluru, 560012, India

***Abstract*—An optical camera-based intrusion classification system (Light Intrusion DeTection systEm named as acronym LITE) for an outdoor setting was recently developed by a superset of the authors. The system classified between human and animal images captured in a side-view manner based on the height. Based on the system and algorithm design, most probably human-crawl would be classified as animal by the LITE. In this paper, classification between human-crawl and animal is addressed. In addition to this work, classification of person with weapon versus person with vehicle is also addressed (referred as person with object) to provide more information about the type of intrusions. A Convolutional Neural Network (CNN) based approach is used to solve the above stated two problems. In the case of "person with object" classification, a study of different CNN architectures was carried out and analysis corresponding to that is presented. In case of human crawl vs animal movement, performance results corresponding to only the best architecture model is provided among the many tried models. Further on, additional insights are provided about the classification using the attention heat maps and t-SNE plots. The test classification accuracies for human-crawl vs animal and person with object classification on the recorded data are close to 95.65% and 90%, respectively. The LITE, having the Odroid C2 (OC2) Single-Board Computer (SBC) with CNN-based classification algorithm for human-crawl versus animal task ported on it, was deployed in an outdoor setting for a real-time deployment. It provided a classification accuracy close to 92%.**

***Index Terms*—Deep learning, Convolutional Neural Networks, Human-crawl vs Animal Classification, Person with Object Classification, Attention Heat Maps**

I. Introduction

a) Motivation: An optical camera-based intrusion classification system (named as LITE) was developed for intrusion detection and classification in an outdoor environment by a superset of the authors in the recent paper [1] (see Subsection II-A for the deployment site). Human and animal classes were classified based on the height information captured from the optical flow and temporal-difference based features generated from the videos. A human crawl would be misclassified by the LITE as an animal class as it extracts the height information for the classification as shown in Fig. 1. In this paper, classification between human-crawl and animal (here only dog data are considered) movement is addressed.

In addition to this work, classification of person with weapon versus person with vehicle is also addressed to provide more information about the type of intrusions. This information is helpful as the long-term goal is to deploy the LITE on a periphery of forest areas where there is a possible movement of poachers with either vehicle(i.e. usually a two wheeler etc.) or a weapon(axe, sickle, saw and long stick etc.).

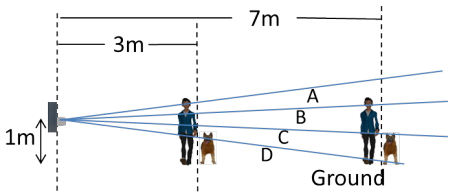


Fig. 1. The height based intrusion classification from the camera by earlier algorithms in LITE.

Hence, a CNN-based approach is explored for human-crawl vs animal movement classification as well as for "person with object" (person with vehicle or weapon) classification. The CNN-based approach has an advantage that a lot of time can be saved which otherwise would be required in feature engineering.

b) Related Work: We list below few papers which are related to the present work. [2] discusses about the handgun detection alarm in videos using deep learning. It provided an F-score of 91.43%. While, [3] discusses the image based vehicle analysis in which it is mainly distinguishing between different vehicles. It gave the best accuracy of 99.50%. The human crawl was detected using pyroelectric infrared sensor in [4] but the paper didn't differentiate between the human crawl and animal movement. Micro-Doppler signals are used for classification between walking, running, creeping, and crawling in [5] and provided 88.97% accuracy for the crawl. The information from the intensity and thermal infrared cameras are fused in [6] to classify human crawl from the other movements (walk and run). It provided 83% recall and 79.6% precision for the human crawl. A classification between different classes using optical camera images is carried out including the adult crawl in [7] and provided 81% of accuracy for the adult crawl. As shown in Table I and [8], different classes including human crawl were classified using different classifiers and 91.89% accuracy was obtained for the human crawl. The present paper differs from existing work in the literature on different fronts. A comparison of present work with existing related work is presented in Table I.

c) Present Work: The primary contributions of the present work are

1) Developing a CNN architecture for the human-crawl vs animal classification. This architecture is implemented on a Single Board Computer (SBC) and the performance is tested in a real-time outdoor deployment setting.
2) Developing different CNN architectures for the people

978-1-5386-5314-2/18/$31.00 ©2018 IEEE

TABLE I
AN OVERVIEW OF SOME RECENT LITERATURE RELATING TO THE CURRENT WORK.

Olmos et al [2] 2017	Videos from the youtube	Hand-gun Detection	F-score of 91.43 %	In our work, person with weapon and person with vehicle are classified and algorithm is implemented on a single board computer
Zhou et al [3] 2016	Optical Camera	Vehicle classification	Best accuracy of 99.50 %	In our work, person with weapon and person with vehicle are classified and algorithm is implemented on a single board computer
Kastek et al [4] 2008	Pyroelectric Infrared sensor	Human crawling detection	NA	In our case, optical camera is used. And, along with detection, classification between human crawl and animal is carried out
Tekeli et al [5] 2016	Micro-Doppler Radar	Classification between walking, running, creeping, and crawling	88.97% accuracy for crawl	In our work, optical camera is used
Zweng et al [6] 2015	Intensity and thermal infrared cameras	Fusing the thermal infrared and stereo depth information	For human crawl, 83.0% Recall and 79.6% precision	In our case, only optical camera is used
Yan et al [7] 2015	Optical Camera	Classifying crawl with different classes	81% accuracy for the adult crawl	In our work, classification accuracy is more though the number of classes are also less
Bogomolov et al [8] 2003	Optical Camera	Classification between vehicle, animal, human, group of people, crawling man and others using template based classifier, Mahalanobis distance based classifier and SVM classifier	91.89% accuracy for crawl	In our case, convolutional neural network is used to reduce the time required for the feature engineering

with object classification and studying the effects and performances of these architectures.

The organization of the paper is as follows. Section II provides infomation related to the experimental setup which includes data collection and data augmentation. Section III provides information regarding the implementation setting for different parameters for CNN architectures. Section IV describes the details of the different combination of CNN architectures developed for the "person with object" classification model, analysis and inferences corresponding to those. Section V provides a description of the CNN model developed for the human-crawl vs animal classification. Section VI describes about the deployment settings and the statistics on the real time classification of human crawl vs animal model. Section VII concludes the paper and states the future work.

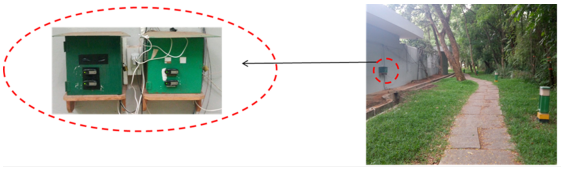


Fig. 2. The snapshots of one of the data-collection site which is also used for deployment as mentioned in Section VI.

(a) (b)

Fig. 3. Snapshot of images captured in a side-view manner at Main Guest House, IISc, for the human-crawl and animal: (a) Human-crawl (b) Dog

II. EXPERIMENTAL SETUP

A. Data Collection

Most of the data corresponding to different classes were collected at Main Guest House (MGH), Indian Institute of Science (IISc) by using the prior deployment settings as in [1] (see Fig. 2). The MGH deployment setting consists of a trail at a distance of around 4m from the LITE and LITE was placed perpendicular to ground at a height of around 1m. The LITE monitors the side-view of intrusions movement along the trail. Additionally, some data corresponding to dog intrusion were also collected at a dog training facility in Kengeri, Bangalore. At Kengeri, dog movements were captured till the distance of 7m. Some data corresponding to human intruder were also collected at the signal processing building of IISc by placing the LITE at a height of 1m from the ground. A typical snapshot images of human-crawl and the dog movement is shown in Figs. 3(a) and 3(b).

B. Data Augmentation

In general, the accuracy of CNN increases with the addition of datapoints. In our case, the number of collected data points were less. Hence, data augmentation technique was used to increase the number of data points for the training. Additionally, the data augmentation process acts as a regularizer, which helps to reduce the generalization error on the real time data as the model learns more variations. The "ImageDataGenerator" function of the Keras [9] was used for the data augmentation. Different parameters used for data-augmentation includes angular-rotation, height-shift-range and width-shift-range of values within 20 degrees, 0.1(fraction of the total height) and 0.1(fraction of the total width), respectively. In the next step, all the images were reduced to the dimension of (150×150). These images were used for training the CNN. The databases and the number of data-points corresponding to different classes are mentioned

in the Table II. Data for validation set and test set are kept separate which are not having the augmented images.

TABLE II
TOTAL NUMBER OF DATA-POINTS WITHIN DIFFERENT DATABASES.

Human-crawl vs Animal Classification Database		
Human-crawl	Animal	Total
3426	3644	6892
Person with Object Classification Database		
Person with weapon	Person with vehicle	Total
1739	2921	4660

III. IMPLEMENTATION DETAILS

The code was implemented in Keras library of python. Following are the default parameters for the CNN architectures used in this paper. If there are deviations from these parameters then those are mentioned in the corresponding architecture when those are described. The number of convolution-filters used in different layers of CNN were of sizes 12, 52 and 60. Each CNN filter was of the size of (3×3). The maxpool operation of default size was (2×2). The ReLU as shown in Eq. 1 was used as an activation function (x is an input to ReLU). It maps the negative values to zero and provides the positive value as it is. The loss function used was categorical cross entropy with rmsprop as an optimizer. Dropout rate of 0.5, was used to reduce overfitting. Softmax was used as a classifier as shown in Fig. 4 and Eq. 2. It provides the probability value for each class as an output and whichever class obtain the maximum probability, the input is declared to belong from that class. The number of epochs was set to 30, which was limited by time and memory constraints. The batch size typically used was 32.

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

$$softmax(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}} \tag{2}$$

Where,

$$z_i = w_i^T x + b \tag{3}$$

Where, w_i is a weight vector, x is an input, b is a bias, i is the class number and N is the total number of classes.

IV. CNN ARCHITECTURES FOR THE PEOPLE WITH OBJECT CLASSIFICATION

The classification accuracies obtained by a CNN is mainly dependent on the choice of its architecture. A description corresponding to different layers of the CNN is provided in [10]. Hence, it is important to choose correct combination of parameters for the CNN to achieve good accuracy. Experimental studies with respect to different combination of parameters for speech recognition task were carried out in [11] and [12]. An experimentation for choosing the correct combination of parameters for the "person with object" classification task was carried out and details corresponding to that are provided further.

A. Number of Convolutional vs Fully Connected Layers

Choosing the correct number of convolutional and fully-connected layers is important to achieve better accuracies. The convolutional layers are used to extract the feature information. The fully connected layers combine the information learned by the convolutional layers for the classification task. Different combinations of convolutional and fully-connected layers were tried out and details corresponding to that are provided in the Table III. In Table III, the notation of $conv3-12$ is used which is described as follows: $conv$- refers to convolution operation, 3 is the size of the filter (i.e. filter of size (3×3), and 12 is the number of filters.

1) Analysis and Inferences: All the analysis and inference are based on our database. It can be seen that $Arch-1$ consisting of only dense layers leads to poor accuracy on the training set, because of the fact that the model has not learned the enough features and it is underfitting (has low training accuracy). In our case, as number of convolution layers increased and number of dense layers decreased to maintain the parameters as constant, an increase in accuracy was observed ($Arch-1$ to $Arch-3$). Addition of one more convolution layer to $Arch-3$ (while keeping the dense layers same) resulted in $Arch-4$, and $Arch-4$ performed better compared to $Arch-3$. This is probably because last convolution layer would be learning more specific feature information. But, $Arch-5$ which is same as $Arch-4$ except that number of units are more in hidden layer performed poor. $Arch-6$ and $Arch-7$ which are though different in number of hidden layers, performed comparatively. $Arch-8$ which has less number of units in the first hidden layer compared to $Arch-7$ performed comparative to $Arch-7$, but $Arch-8$ has a gain as number of parameters are less. $Arch-9$ and $Arch-10$ has same 12 number of filters in all three convolution layers. In general in our case, the architecture which are having more number of filters in the later convolution layer perform better compared to architectures which are having same number of filters in all convolution layers.

Among above mentioned 10 Architectures, $Arch-7$ (see Fig. 4) resulted in better accuracy on the validation set, hence it was considered for the further exploration and analysis.

B. Number of Units in the Fully Connected Layers

Number of units are varied for different dense layers in $Arch-7$ to generate new architectures as described in Table IV. It is observed that the configuration which has the dense layers arranged in such a manner that the number of units in the dense layers are in the descending order (i.e. $Arch-11$) provide by far the consistent results across train, test and validation sets. The $Arch-11$ consists of dense layer of size 256 followed by dense layers of size 128, 64 and 2. Adversely, the $Arch-15$ in which the dense layers are represented in the reverse order of the architecture $Arch-11$, it can be observed that the test accuracy is less. It can be seen that there is a jump from 64 units to 2 units in $Arch-11$, while there is jump from 256 units 2 units in $Arch-15$.

C. Experimentation with Pooling Type and Pooling Size

The pooling layers pools the statistical information from the small neighborhoods and generally construct a smaller

TABLE III
A PERFORMANCE COMPARISON ACROSS DIFFERENT ARCHITECTURES AS A FUNCTION OF NUMBER OF CONVOLUTIONAL AND FULLY CONNECTED LAYERS. THE (CONVM-N) MEANS N FILTERS OF SIZE $M \times M$.

Arch.	C1	C2	C3	No. of units in H1	No. of units in H2	No. of units in H3	No. of units in H4	No. of parameters	Train. acc. (%)	Val. acc. (%)	Test. acc. (%)
1	0	0	0	256	128	64	2	17321088	68.0	46.88	46.67
2	conv3-12	0	0		256	128	2	16855620	99.34	84.18	86.15
3	conv3-12	conv3-52	0	0	0	256	2	17258804	98.78	87.11	87.52
4	conv3-12	conv3-52	conv3-60	0	0	256	2	4473572	99.06	90.82	90.25
5	conv3-12	conv3-52	conv3-60	0	0	1024	2	17792228	68.09	48.99	46.67
6	conv3-12	conv3-52	conv3-60	0	1024	64	2	17855844	96.30	89.26	91.45
7	conv3-12	conv3-52	conv3-60	1024	128	64	2	17929572	93.37	91.41	90.08
8	conv3-12	conv3-52	conv3-60	256	128	64	2	4514148	95.28	90.04	90.08
9	conv3-12	conv3-12	conv3-12	256	128	64	2	931812	97.41	88.87	88.88
10	conv3-12	conv3-12	conv3-12	0	0	64	2	224996	95.44	90.23	88.37

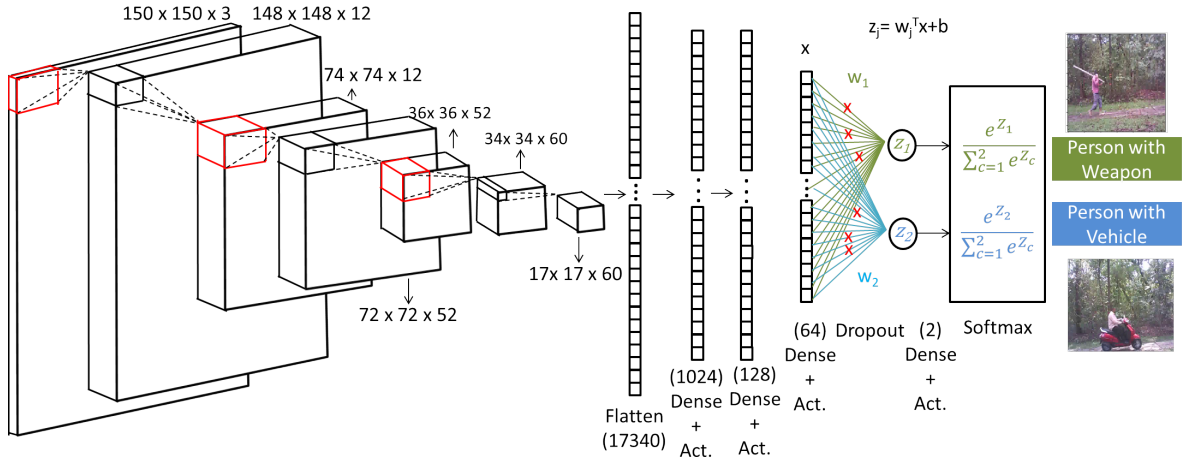


Fig. 4. The $Arch-7$, complete architecture of convolutional neural network used as a reference architecture.

TABLE IV
A COMPARISON ACROSS DIFFERENT ARCHITECTURES (ADAPTION OF $Arch-7$) AS A FUNCTION OF NUMBER OF HIDDEN UNITS.

Arch.	# units (H1)	# units (H2)	# units (H3)	# units (H4)	Train. acc. (%)	Val. acc. (%)	Test. acc. (%)
11	256	128	64	2	95.28	90.04	90.08
12	150	150	150	2	96.86	90.23	81.53
13	128	128	128	2	95.21	90.23	46.67
14	0	128	64	2	92.60	88.28	88.20
15	64	128	256	2	93.57	91.21	46.67

size image from the extracted statistical information. There are different pooling options to extract the statistical information like max-pooling and average-pooling.

In the Table V, experimentation with different pooling types and pooling sizes is carried out in the convolutional layers of the $Arch-7$. It is observed that the maxpool performs better compared to average pooling. Because the current work is focused on the object detection, and the maxpooling operations help extracting edges as features which is critical in distinguishing between the objects. On the other hand, average pooling smoothens the image, which is not helpful in object detection as it can be inferred from the results as shown in Table. V. The performance results across train, validation and test sets for the maxpool operations are consistent($> 90\%$) for maximum number of cases.

TABLE V
A PERFORMANCE COMPARISON ACROSS DIFFERENT ARCHITECTURES (ADAPTION OF $Arch-7$) AS A FUNCTION OF POOLING TYPE AND POOLING SIZE.

Arch.	Pooling type	Pooling size	Train. acc.(%)	Val. acc.(%)	Test. acc.(%)
16	Maxpool	2	93.37	91.41	90.08
17	Maxpool	3	91.99	89.26	92.99
18	Maxpool	4	91.89	88.67	90.25
19	Avgpool	2	94.98	88.48	89.05
20	Avgpool	3	93.60	89.65	83.76
21	Avgpool	4	89.05	84.77	84.27

D. Experimentation with Optimizers

Table VI provides the $Arch-22$ and $Arch-23$ which are adaption of $Arch-7$ in terms of different optimizers

used such as rmsprop and adam. The rmsprop performs better compared to adam in this case as shown in Table. VI.

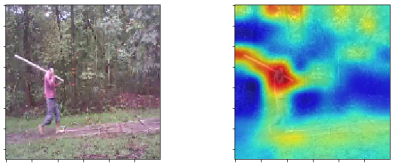
TABLE VI
A PERFORMANCE COMPARISON ACROSS DIFFERENT ARCHITECTURES (ADAPTION OF $Arch-7$) AS A FUNCTION OF OPTIMIZER USED.

Arch.	Optimizer used	Train. acc.(%)	Val. acc.(%)	Test. acc.(%)
22	adam	95.73	86.72	89.40
23	rmsprop	93.37	91.41	90.08

.

Attention Heat Maps: The attention heat maps as described in [13] are used to visualize the underlying features which are differentiating between the two classes. The Keras-vis tool is used for the purpose of visualizing the heat map.

The heat map for the weapon model has very high intensity region covering the head, arms and the carrying weapon, in this case a stick, as depicted in the Fig. 5(b). While, the heatmap for vehicle is more concentrated on the wheels and body of the vehicle as shown in Fig. 6(b).



(a) (b)

Fig. 5. Example images from the class corresponding to human carrying a weapon. In (a), image of a human carrying a weapon in an outdoor setting is shown. In (b), the attention heat map corresponding to same image is presented.

(a) (b)

Fig. 6. Example images belonging to class corresponding to human with vehicle. In (a), an image of man driving a vehicle in an outdoor setting is shown. In (b), the attention heat map of the same image is presented.

V. CNN ARCHITECTURE FOR THE HUMAN CRAWL VS ANIMAL CLASSIFICATION

The architecture proposed in the present work for human crawl vs animal classification is similar to LeNet by [14], but with additional dense layers. The baseline architecture used in the current work for the crawl vs animal classification is an adaption of $Arch-6$ of Table III. Addition to $Arch-6$, $l2$ regularizer of penalty 0.1 is also employed to reduce the overfitting. The training and the validation accuracies are provided in the Table VII . And, the results corresponding to the test database are shown in the Table VIII which shows an accuracy close to 96%.

A. Heat maps for human-crawl vs animal

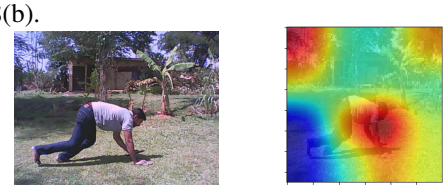
In case of human-crawl vs animal, the distinction between the two classes can be seen from the heat-map of some of the images. The heat map for the human-crawl is more intense and concentrated towards the arms and the head region as shown in Fig. 7(b). On the contrary, heat map for the dog is not intense, distributed along most of the body as depicted in Fig. 8(b).

TABLE VII
THE CLASSIFICATION ACCURACIES CORRESPONDING TO HUMAN CRAWL VS ANIMAL CLASSIFICATION

Classifier	Train. Acc	Val Acc.
CNN	96.65	98.60

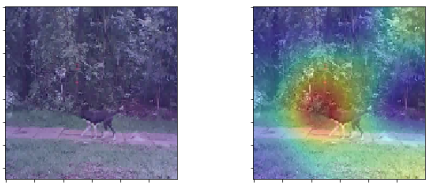
TABLE VIII
THE CLASSIFICATION ACCURACIES ON TEST DATABASE OF HUMAN CRAWL VS ANIMAL CLASSIFICATION

Class	Total Examples	Correct Classifications
Human Crawl	520	483
Animal	378	376



(a) (b)

Fig. 7. Images corresponding to the human crawl class. In (a), human crawl image in an outdoor setting is shown. In (b), attention heat map of the same image is depicted.



(a) (b)

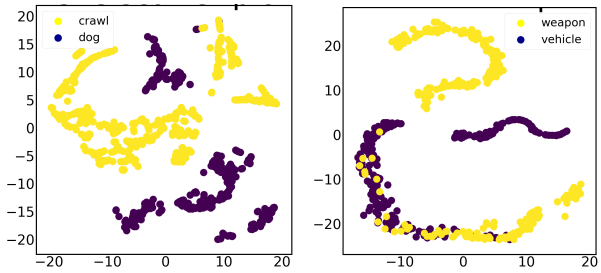
Fig. 8. Images corresponding to the dog class. In (a), dog image is shown in a typical outdoor setting. In (b), attention heat map of the same image is depicted.

B. t-SNE plots

The t-SNE plot as discussed in [15] is useful to analyze and visualize the class separability distribution of features by mapping them to a lower dimensional space. Fig. 9(a) shows the t-SNE plot for the human-crawl vs animal on the penultimate layer which is of 64 dimension and correspondingly for person with object classification in Fig. 9(b). It has been observed that classes are having considerable non-overlap region for the human crawl versus animal model. On the contrary, the human weapon versus the human vehicle model to certain extent has a considerable non-overlap region, but not as distingusihable as in the case of the human crawl vs animal model.

VI. DEPLOYMENT AT IISC

Optical camera system was developed [1] which used an SVM as a classifier to distinguish between human, animal and vegetative motion. Once the SVM classifier detects an



(a) Human-crawl vs animal (b) Human with weapon vs human with vehicle

Fig. 9. t-SNE plots of two models. First being the human-crawl vs dog and second being human with weapon vs human with vehicle. (a) A t-SNE plot for human crawl vs animal dataset. The yellow and purple markers represent data-points corresponding to human crawl and dog classes, respectively. (b) A t-SNE plot for human with weapon vs human with vehicle model. The yellow and purple markers represent data-points corresponding to weapon and vehicle classes, respectively.

animal, twenty frames of the detected animal intrusion are stored in a frame buffer in First In First Out (FIFO) manner.

Out of these only thirteen frames (i.e. frame No 2 to 14) are read by the CNN classifier i.e. trained on $Arch - 6$ for human crawl versus the animal. The purpose of using thirteen frames is found out by experimentation considering that there is more likelihood of presence of intrusion in these frames. The CNN classifier reads those images and further classifies between human-crawl and animal. The final decision as human-crawl or animal is decided based on the majority vote on the decision over those 13 images. This model is deployed on an Odroid C2 (a single board computer) to test the performance in a real time outdoor setting. The deployment site settings can be as observed in the Fig. 2 where movement of human crawl and dog model were observed on the trail. Crawl movements were carried out by the team members to test the efficacy of the system as human-crawl movements are not generally observed in this setting.

A. Field Testing Results

Table IX summarizes the performance results of the field testing corresponding to human-crawl versus animal model from 29th October, 2017 to 04th November, 2017. The camera system is planned to be used for only the day-time data. Hence, performance statistics are generated for only the day-time.

An average accuracy for human-crawl is 91.66% with total human-crawl intrusions as 48. And, accuracy for the dog class is 94.11% with a total dog intrusions as 17. Since, the movement of the dog on the trail is a rarity, the model has to be deployed significantly for many days to get more intrusions. The classification results on the real time deployment settings are presented in the Table IX.

TABLE IX
ACCURACY OF THE CNN ARCHITECTURE. (FOR 7 DAYS)

Class Type	Total Intrusions	Correct Detections	Mis-classifications
Human-Crawl	48	44	4
Dog	17	16	1

VII. CONCLUSION AND FUTURE WORK

In this paper, a study and analysis was carried out for classification between human with weapon and human with vehicle in an outdoor setting using different CNN architectures. The best model provided an accuracy close to 90%. Additionally, a CNN architecture was also designed for human-crawl versus animal classification which provided an accuracy of around 96%. This model was also implemented on an Odroid C2 SBC and was deployed in a real time outdoor setting for a duration of 7 days which provided an accuracy close to 92%. The Inferences of attention heat maps show distinguishable features for both the models and t-SNE plots show good separability between the two classes.

Future work will include extending the deployment to human with weapon versus human with vehicle model for real time classification. Also, a complete system need to developed which doesn't depend on earlier classification algorithms based on the SVM. The future work needs to address the development of multi-class robust CNN model which should be able to classify between clutter, human, animal, human-crawl, human with weapon, and human with vehicle classes. Hence, a 6 class robust CNN model need to be developed which can classify between all the above classes.

REFERENCES

[1] T. Choubisa *et al.*, "An optical-camera complement to a PIR sensor array for intrusion detection and classification in an outdoor environment," *submitted to SenseApp 2017*. [Online]. Available: https://goo.gl/ZPMMXA

[2] R. Olmos, S. Tabik, and F. Herrera, "Automatic handgun detection alarm in videos using deep learning," *Neurocomputing*, 2017.

[3] Y. Zhou *et al.*, "Image-based vehicle analysis using deep neural network: A systematic study," in *Digital Signal Processing (DSP), 2016 IEEE International Conference on*.

[4] M. Kastek, T. Sosnowski, *et al.*, "Long-range pir detector used for detection of crawling people," in *Proc. of SPIE*, vol. 7113, 2008.

[5] B. Tekeli, S. Z. Gurbuz, and M. Yuksel, "Information-theoretic feature selection for human micro-doppler signature classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 54, no. 5, pp. 2749–2762, 2016.

[6] A. Zweng, C. Beleznai, and C. Sulzbachner, "Reliable intruder detection using combined modalities of intensity, thermal infrared and stereo depth," in *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*. IEEE, 2015, pp. 1–6.

[7] Y. Yan, C. Xu, D. Cai, and J. Corso, "Weakly supervised actor-action segmentation via robust multi-task ranking," *language*, vol. 48, p. 61.

[8] Y. Bogomolov *et al.*, "Classification of moving targets based on motion and appearance." in *BMVC*, 2003, pp. 1–10.

[9] ImageDataGenerator, "Imagedatagenerator," [accessed 05-Nov-2017]. [Online]. Available: https://keras.io/preprocessing/image/

[10] Y.-J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning-based crack damage detection using convolutional neural networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 5, pp. 361–378, 2017.

[11] T. N. Sainath *et al.*, "Deep convolutional neural networks for lvcsr," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8614–8618.

[12] ——, "Improvements to deep convolutional neural networks for lvcsr," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 315–320.

[13] B. Zhou *et al.*, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[14] Y. LeCun *et al.*, "Lenet-5, convolutional neural networks," *URL: http://yann. lecun. com/exdb/lenet*, 2015.

[15] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.