HMD Vision-based Teleoperating UGV and UAV for Hostile Environment using Deep Learning

Abhishek Sawarkar¹, Vishal Chaudhari, Rahul Chavan, Varun Zope, Akshay Budale and Faruk Kazi

Abstract—The necessity of maintaining a robust antiterrorist task force has become imperative in recent times with resurgence of rogue element in the society. A well equipped combat force warrants the safety and security of citizens and the integrity of the sovereign state. In this paper we propose a novel teleoperating robot which can play a major role in combat, rescue and reconnaissance missions by substantially reducing loss of human soldiers in such hostile environments. The proposed robotic solution consists of an unmanned ground vehicle equipped with an IP camera visual system broadcasting real-time video data to a remote cloud server. With the advancement in machine learning algorithms in the field of computer vision, we incorporate state of the art deep convolutional neural networks to identify and predict individuals with malevolent intent. The classification is performed on every frame of the video stream by the trained network in the cloud server. The predicted output of the network is overlaid on the video stream with specific colour marks and prediction percentage. Finally the data is resized into half-side by side format and streamed to the head mount display worn by the human controller which facilitates first person view of the scenario. The ground vehicle is also coupled with an unmanned aerial vehicle for aerial surveillance. The proposed scheme is an assistive system and the final decision evidently lies with the human handler.

I. INTRODUCTION

Governments around the world are investing billions of dollars to safeguard their citizen and secure their borders to maintain the territorial integrity and sovereignty of the nation. The risk of a terrorist attack can never be eliminated but prudent steps can be taken to reduce this risk. However, in certain cases due to the lapse in security measures, few rogue elements threaten to exploit human fears to help achieve their goals. They subjugate innocent civilians to get their demands fulfilled. With the advancements in technologies and advent of intelligent systems, these rogue elements can be neutralized with minimum human casualties. Intelligent robots can be used for combat or spy operations or for different purposes ranging from mine detection, surveillance, logistics and rescue operations to reconnaissance and support, communications infrastructure, forward-deployed offensive operations, and even as tactical decoys [1]. However, it is unreliable to give complete control to autonomous systems to handle such dangerous situations. Therefore, a semi-autonomous system in which computer intelligence assists the human operator will provide a better outcome. A system involving innate human intelligence coupled with robot's ability to process huge amounts of data can function effectively in such scenarios. The final call to make decision

All authors are with Department of Electrical Engineering, Veermata Jijabai Technological Institute (VJTI), Mumbai, India.

1 sawarkarabhi@gmail.com

will evidently remain with the human while the machine could advocate its results based on the incoming data.

Such systems are being developed by various defence agencies like DARPA's LAGR program [2] which focused on perception based off-road navigation in UGV's. Foster-Miller's TALON system is a remotely operated vehicle designed for missions ranging from combat to reconnaissance. They are working on incorporating virtual reality vision for their future vehicle called the Modular Advanced Armed Robotic System (MAARS) [3]. The US marines use their Gladiator Tactical Unmanned Ground Vehicle (TUGV) to minimize risks and eliminate threats during conflict. This vehicle is light weight and can be easily transported and deployed strategically for missions. The Indian Defence Research and Development Organization (DRDO) developed a fully automated UGV named Daksh for handling and destroying hazardous objects safely. These vehicles can be controlled remotely ensuring the safety of the operator and minimizing human loss.

Tele-operation capabilities, or the ability for an operator to manipulate and control a robot remotely from a safe location through a radio link offers the possibility of controlling the robot in hostile environment with no human casualties. These capabilities reduce or remove operator's risk in highly stressful and dangerous environment. The human operator based on a remote location, away from the danger, receives the data wirelessly from the robot in real-time. This data is processed and presented upfront such that it does not overload the operator. proposes a robotic solution for handling terrorist situation in conjunction with human assistance. The machine can be used for first assault in hostile environment. The system comprises of tele-operated unmanned ground vehicle (UGV) and an unmanned aerial vehicle (UAV) that relays real-time information to the human controller wirelessly. The two robots provide multiple perspective of the environment to the controller [4]. The UGV-UAV pair is inspired by the work done by Cantelli et al. [5] for surveying operation. The system can classify terrorists from hostages by using a trained Deep Convolutional Neural Network. The control and operation of the machine is in the human controller's hands. By converting the video information in virtual reality (VR), the human controller can experience an immersive first person view of the situation [6]. The person in consideration for classification is underscored by a bounding box with appropriate tag which is augmented on the video stream. The VR video system assists the operator to neutralize threat by detecting armed personnel in real-time.

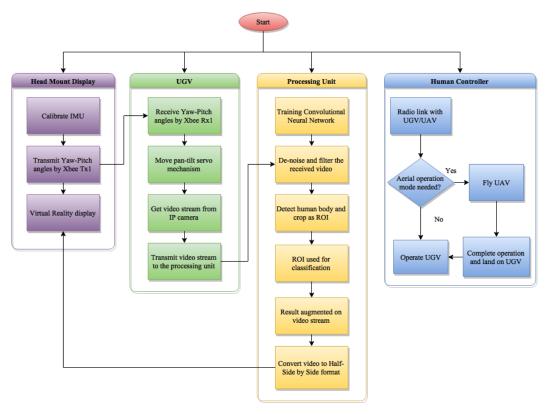


Fig. 1: Flowchart of proposed system

II. HARDWARE

The hardware comprises of two robots, the Unmanned Ground Vehicle (UGV) for land based operation an the Unmanned Aerial Vehicle (UAV) for air surveillance.

A. UGV

An Unmanned Ground Vehicle (UGV) is a land based vehicle without a human on-board which can be used to perform civilian or dangerous military tasks. These vehicles are used to replace humans to work in perilous conditions like bomb defusal, dilapidated nuclear reactors, surveillance and tactical situations etc.

The UGV used in this project has a six wheel differential drive mechanism that enables navigation in off-road and rugged terrain. The wheel mechanism is inspired by rockerbogie system [7] with additional wheels at the front and back of UGV. This mechanism facilitates the UGV to climb steps and steep slopes of maximum 50 degrees inclination. The UGV is equipped with a servo pan-tilt mechanism which holds an IP camera. The entire pan tilt mechanism is mounted on the UGV base for stabilizing the video stream. This servo mechanism is controlled by head tracking inertial measurement unit(IMU) placed on the head mount VR goggle worn by the human controller. IP camera on the UGV wirelessly streams the reltime video input to a remote processing unit for classification and teleoperation.

B. UAV

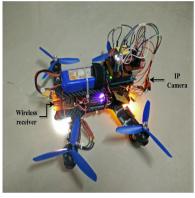
The UAV stands for Unmanned Aerial Vehicle which is an aircraft without human on-board. UAV's can be remotely controlled by a human controller or it can fly autonomously based on pre-programmed flight plans. The UAV used in this project is remotely controlled first person view (FPV) type quad-rotor. This quad-rotor is mounted on the UGV base. Areas where the UGV fails to travel due to motion constraints, the operator can switch to aerial mode and fly UAV to complete the operation and land the UAV back on UGV. The Quadrotor frame is 'X' shaped of dimension 250mm measured diagonally from motor shaft-to-shaft. The carbon fiber body frame makes the UAV durable and light weight. A FPV camera relays the video stream when the remotely located human controller switches to the aerial mode. While the UGV could be used for combat scenarios providing attack capabilities in hostile environments, the UAV proposed in this project can only be used for reconnaissance and surveillance missions to provide vital information.

III. SOFTWARE

A. Image Processing

The video stream received from the IP camera mounted on the UGV is pre-processed and analyzed by the remotely located processing unit. The video data is blurred to remove noise by Gaussian smoothing using a Gaussian kernel [8]. Gaussian function transforms each pixel's original value to weighted average of that pixel's neighbourhood. The 2D





(b) UAV

Fig. 2: Unmanned Ground Vehicle and Unmanned Aerial Vehicle

Gaussian function is given by:

$$G_0(x,y) = \frac{1}{2\pi\sigma^2} e^{\left(\frac{-(x-\mu_x)^2}{2\sigma_x^2} + \frac{-(y-\mu_y)^2}{2\sigma_y^2}\right)}$$
(1)

 μ_x = mean in horizontal axis

 μ_y = mean in vertical axis

 σ^2 = variance

After initial de-noising by the Gaussian filter, the video stream is processed to detect a human in the frame. Two algorithms are evaluated to find a person in the frame -

1) HOG + SVM: The Histogram of Oriented Gradients (HOG) descriptor for object recognition provides excellent performance for human detection. The local object appearance and shape can often be distinguished by the distribution of local intensity gradients or edge direction as described by Dalal et al. [9]. The HOG descriptor identifies the feature set of the human figure based on the samples of positive and negative images provided during training. While testing the input images are classified as positive or negative by a linear Support Vector Machine (SVM). The frame containing a human figure is classified as positive else it is classified as negative.

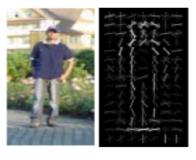


Fig. 3: Sample image of a person and its HoG visualization

The major disadvantage of the above method is that the classifier is used as a sliding window detector on an input image. The sliding window technique for classification of entire human frame is computationally expensive. Also, to detect all instances of humans at multiple scales, the sliding

window has to run at multiple scales to form pyramid of detector responses. Any occlusion of the human results in loss of detection. The HOG+SVM algorithm used in this project was implemented using the OpenCV library to find a person in the video stream. The project demands a real-time performance as the robot will be used in hostile environment. However, the CPU implementation on Intel i3 with 4GB RAM and 2.4GHz processor was not realtime. An input video stream at 30 fps from IP camera on the UGV resulted in 8 fps output. The GPU implementation on Nvidia GPU GeForce GT 755M and 2GB RAM was better than the CPU with output at 20 fps. Another algorithm proposed by Viola and Jones [10] was used to detect a face and the person in realtime.

2) Haar feature based Cascade Classifier: This algorithm incorporates different kind of feature set based on Haar wavelets instead of the usual image intensities as proposed by Papageorgiou et al. [11]. A Haar-like feature stated in [12] considers adjacent rectangular regions at a specific location in a detection window. It sums up the pixel intensities in each region and calculates the difference between these sums. This difference is then used to categorize subsections of an image. These Haar-like features can be calculated in constant time which make it faster than other algorithms.

The main advantage of Haar based cascade classifier over other algorithms is the realtime detection of an object, in this case detection of a person. The algorithm proposed by Viola and Jones [10] uses several simpler classifiers that are applied subsequently until at some stage the image is rejected or all stages are passed. If all stages are passed, the object of interest is detected and marked as the Region of interest (ROI). Based on human body geometry, the entire human anatomy is detected by the aforementioned algorithms and selected as the ROI. The ROI is cropped and resized to resolution of 227x227 pixels which is then forwarded to a trained Deep Network for classification of the person as probable terrorist.

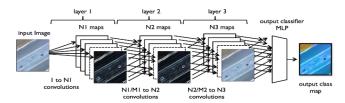


Fig. 4: General architecture of Convolutional Neural Network

B. Deep learning and classification

A person detected as the ROI using human detection algorithm can be classified as a potential terrorist by assessing the features of a terrorist. These features may include weapons like a gun or assault rifle, facial emotions, walking patterns etc. This project considers a gun or assault rifle for classifying the person holding the weapon as a probable terrorist.

A Deep Learning neural network called the Convolutional Neural Network (CNN) is trained to identify human with guns like assault rifles, revolvers, etc. Deep learning aims at learning feature hierarchies wherein features in higher levels of the hierarchy are formed from lower level features extracted in previous stages [13]. The salient advantage of CNN over other neural architectures is that the neurons in the higher layers have local connectivity with the neurons in the previous layer which reduces the number of connections and the weights to be trained. In order to improve generalization and reduce the number of hyper-parameters, a convolution operation on small region of the input image is performed. CNN utilizes 'shared weight' in the convolutional layers that reduces space complexity and improves performance.

The CNN architecture used in this project is based on Inception architecture [14] formulated by GoogLeNet team which participated in ILSVRC14 competition which is based on the LeNet-5 network developed by LeCun et al. [15]. The network is 22 layers deep when counting only layers with parameters. The GoogLeNet team had trained the network on ImageNet dataset which comprises of 1.2 million images for training, 50,000 for validation and 100,000 for testing ranging across 1000 different categories. The final layer of the network is called the 'Softmax loss function' which is a linear layer that predicts the output out of 1000 classes. The 1000 different categories included both animate and inanimate objects. The network was trained to detect objects like car, hammer, gun, revolver etc for ImageNet Large Scale Visual Recognition Challenge 2015 [16].

In this project, we utilize this pre-trained model to learn to classify particular object in the input image. We utilize 'Transfer Learning' [17] technique in Caffe Deep Learning Framework [18] to fine-tune the inception network in-order to detect probable terrorists as training the entire network for classification from scratch was not feasible given the limited data and hardware resources. The CNN was fine tuned to classify only 8 different classes instead of 1000 and the last layer of the network was changed in the model. The weights for the new last layer were randomly initialized and the

model was re-trained on the new dataset comprising of different types of hand held weapons like assault rifle, revolver, pistol etc. The custom dataset constituted positive images of firearms and arbitrary negative samples of images not containing a weapon. Deep Convolutional Neural Network is capable of achieving better results for image classification than conventional algorithms. The network could achieve an error rate of 40% on the custom terrorist dataset with the test inputs from highly noisy video stream. A bounding box of specific colour and prediction probability of the network are overlaid on the video stream to assist the human controller.

If the human in the input is classified as a civilian, the bounding box is shown in 'Green' colour. However, if the network classifies the person as a probable terrorist, then the bounding box is displayed in 'Red'. Subsequently, the video stream is formatted and streamed to head mount display (VR goggles).



Fig. 5: Head mount display using VR goggle

C. Head Mount Display

Head mount displays (HMD) are increasingly being used in cockpits of modern helicopters and air-crafts to assist soldiers. These devices can display important tactical information along with the real scene that provides an effect of total immersion and a first person's perspective. The immersive vision helps the soldier to get a better understanding of the situation and control the remote robot with ease.

The HMD used in this project uses a smartphone virtual reality (VR) goggle that receives formatted video stream from the processing unit. The output frames given by the CNN are augmented with appropriate bounding boxes based on the classification results of CNN. These augmented frames are split into Half side-by-side format with each side having resolution of 950x1000. The frames are wirelessly streamed to the smartphone connected in the Ad-Hoc connection and placed in the VR google.

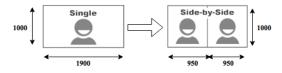


Fig. 6: Half Side-by-Side conversion

The human controller wearing the goggle gets a first person view of environment as seen by the UGV and UAV.

The goggle is equipped with motion sensors (IMU) to track human controller's head movements which are calibrated at initialization. An accelerometer on the goggle records the pitch angle while a magnetometer notes the yaw orientation of the head movements. These angles are passed through a low pass filter to remove jitters and sharp reading changes. A Xbee transmitter relays the head tracking data to a Xbee receiver mounted on the UGV that controls the pan-tilt mechanism. As a result, the IP camera on the mechanism rotates as per the head movements of the human controller in real-time.

IV. IMPLEMENTATION

The operator based at a remote location initiates the UGV via radio link. At the same time, the head mount display is calibrated by the operator to track his head movements. The head motion data is relayed to the UGV through XBee Rx-Tx pair based on IEEE 802.15.4 networking protocol. The UGV controls the motion of pan-tilt mechanism based on this data. Another encrypted radio link carries the visual input data of resolution 1900x1000 from an IP camera mounted on the pan-tilt mechanism on the UGV to the head mount display of the operator. Video data is pre-processed and before passing the visual data to the operator, a cloud server running a trained convolutional neural network scans whether the incoming data contains a probable terrorist or not.

A. Case I - Civilian



Fig. 7: As the detected human is unarmed, CNN classifies it as civilian and draws green bounding box around it. It also shows a confidence value (in this case 0), which is percent probability of the person being Terrorist.

B. Case II - Armed personnel



Fig. 8: As the detected human is armed, CNN classifies it as probable terrorist and draws red bounding box around it. The number indicates percent probability of the person being a terrorist

It classifies a person without a gun as a civilian with probability of being a terrorist close to null. A probable terrorist is a human with a gun in our model. The network presents it's classification with a confidence value as viewed at the top of the bounding box and stream is passed to VR formatter. The formatter resizes frames from resolution 1900x1000 into half-side by side configuration with each side of resolution 950x1000. With the head mount display, the operator gets an immersive real-time view of the environment seen by the robot. The situations wherein the UGV cannot operate or it's vision system is restricted, the operator can switch over to aerial mode and operate the UAV to get an aerial perspective of the situation.

V. CONCLUSION AND FUTURE WORK

In this paper we have presented a robotic solution for handling hostile situations thereby reducing casualties and loss of human life. The system comprises of both hardware and software entities to work in such environments. A pair of tele-operating UGV and UAV is proposed which is equipped with wireless vision sensor. Algorithms for human detection namely HOG+SVM and Haar Cascade are presented and a novel technique involving state of the art deep learning architecture for terrorist classification is exhibited for maintaining safety and security and use in hostile environment. The paper also underscores the advantages of head mount display in proposed tele-operating systems.

Currently the system identifies person with a gun as a threat. However, it can be designed to differentiate between an armed serviceman and an armed unknown person. This capability will eventually help in deploying the robot accompanied by task force to counter hostile situations. The face detection technique can be used to detect the person in consideration with pre-compiled database. A match will provide valuable information in such crisis.

The system can be used in different scenarios like disaster response, rescue missions, operation in hazardous radioactive environment etc. The deep convolutional network can be trained to identify specific entities like humans surviving in disaster hit areas, dangerous chemicals in decrepit chemical and radioactive plants, biological samples in such sites etc. A swarm of tele-operating robots, either fully autonomous or semi-autonomous could evidently prove beneficial in mitigating human loss in hostile situations.

ACKNOWLEDGMENT

The work presented in this paper was generously funded by the Centre of Excellence (CoE) in Complex and Non-Linear Dynamical Systems (CNDS) at VJTI.

REFERENCES

- [1] J. Khurshid and H. Bing-rong, "Military robots a glimpse from today and tomorrow," in *Control, Automation, Robotics and Vision Conference*, 2004. ICARCV 2004 8th, vol. 1, Dec 2004, pp. 771–777 Vol. 1.
- [2] L. D. Jackel, E. Krotkov, M. Perschbacher, J. Pippine, and C. Sullivan, "The darpa lagr program: Goals, challenges, methodology, and phase i results," *Journal of Field Robotics*, vol. 23, no. 11-12, pp. 945–973, 2006.

- [3] G. E. Marchant, B. Allenby, R. Arkin, E. T. Barrett, J. Borenstein, L. M. Gaudet, O. Kittrie, P. Lin, G. R. Lucas, R. O'Meara, et al., "International governance of autonomous military robots," Colum. Sci. & Tech. L. Rev., vol. 12, p. 272, 2011.
- [4] J. Y. Chen, "Uav-guided navigation for ground robot tele-operation in a military reconnaissance environment," *Ergonomics*, vol. 53, no. 8, pp. 940–950, 2010.
- [5] L. Cantelli, M. Mangiameli, C. D. Melita, and G. Muscato, "Uav/ugv cooperation for surveying operations in humanitarian demining," in 2013 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), Oct 2013, pp. 1–6.
- [6] R. T. Azuma, "A survey of augmented reality," *Presence: Teleoperators and virtual environments*, vol. 6, no. 4, pp. 355–385, 1997.
- [7] K. Yoshida and H. Hamano, "Motion dynamics of a rover with slip-based traction model," in *Robotics and Automation*, 2002. Proceedings. ICRA'02. IEEE International Conference on, vol. 3. IEEE, 2002, pp. 3155–3160.
- [8] H. J. Blinchikoff and A. I. Zverev, Filtering in the time and frequency domains. Krieger Publishing Co., Inc., 1986.
- [9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1. IEEE, 2005, pp. 886–893.
- [10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition*, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1. IEEE, 2001, pp. I–511.
- [11] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Computer vision*, 1998. sixth international conference on. IEEE, 1998, pp. 555–562.
- [12] R. Lienhart, A. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," in *Pattern Recognition*. Springer, 2003, pp. 297–304.
- [13] Y. Bengio, "Learning deep architectures for ai," Foundations and trends in Machine Learning, vol. 2, no. 1, pp. 1–127, 2009.

- [14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
- [15] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [17] S. J. Pan and Q. Yang, "A survey on transfer learning," Knowledge and Data Engineering, IEEE Transactions on, vol. 22, no. 10, pp. 1345–1359, 2010.
- [18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.