[[1]](#footnote-1)

Leading Athlete following UAV Using Transfer Learning (SSD\_MobileNet\_V1)

Shanmukha Sai Sumanth Yenneti, Chaitanya Chowdary P

*Abstract*—In this paper we consider developing vision-based control algorithms for unmanned aerial vehicles (UAVs), so that a UAV can follow a human athlete autonomously. It can be used as a sports coverage camera and can move anywhere and be mainly used on a racetrack where it is hard to capture the racers and every movement they make. This task involves two sub tasks 1) To detect and mark the leading athlete and 2)To navigate the UAV in such a way that it follows this marked athlete and continuously changes its position with respect to the person which has been marked. Person detection using transfer learning is applied to support our proposed method. Transfer learning is a method of machine learning where a model which has been trained for a particular task can be reused to solve similar tasks to skip the hassle of training especially when there are no resources for computing or storage. Since detecting the person in the video is one of the important tasks while tracking an athlete, we need a model which gives quick and reliable outputs. We will use the person detection on live video and one of the most popular models for this purpose is MobileNet object detection with convolutional neural network-based characteristics because they have high accuracy in analysis. A control algorithm uses the bounding box position to change the UAV flight parameters in order to keep the user in the field of view center. The goal of this paper is to introduce transfer learning for many day-to-day applications of drones, which will make the development and deployment of a drone/ UAV quicker.

*Index Terms*—Bounding boxes, Convolutional neural networks, MobileNet, UAV.

# INTRODUCTION

R

ecently, human monitoring and object detection are seen as the main factors for smart-city surveillance, which can be used to improve digital society security. For several research areas, an unmanned aerial vehicle (UAV), called a drone, has been developed. It can be operated by a person, or a system built to fly the task automatically. Such as military services, disaster relief services, forestry, and transport. This will contribute to the development of the UAV smart system, which is easy to monitor. UAVs that are programmable for a user to operate. That has lots of interesting applications. For example, a UAV could track an athlete, acting as a "personal camera guy." The resulting video could be used for entertainment (a football game filmed from the viewpoint of the quarterback), analyzing performance by athletes (recording the dribbles and field position of a soccer player), etc. A drone may be flying over a first responder or fireman in search and rescue or firefighting, offering an enhanced view of the scene. It may be trailing an elderly person in the sense of assisted living, creating an alarm if the person falls. Finally, a drone may be escorting a child to or from school in a child safety scenario**.** As drones decrease in size, many of these applications, e.g. by fly-sized UAVs, could be introduced almost seamlessly. Currently, only tracking user coordinates with GPS and mobile phones helps the individual to follow. In addition to the well- known unreliability of this solution indoors and in disaster relief scenarios, the above applications involve precise control of drone and subject relative positions being monitored. Often the drone must be kept in front, often behind and sometimes above. In each case, you may need to remain aligned directly with the subject, e.g. immediately above the firefighter, or at an angle, e.g.45 above the soccer player, and monitor the subject from far or near. For instance, the "personal cameraman" should remain close to the technique of recording soccer dribbling, and far from recording player positioning. Such precision in controlling its location allows the UAV to locate the user precisely in the scene, to understand how the user is confronted etc.

Although these objectives could, in theory, be accomplished by equipping drones with computer vision, previous robotic vision work emphasized autonomy, namely autonomous navigation based on simultaneous localization and mapping , visual odometry, and obstacle avoidance; However, these are not major criteria for the above applications where, rather than full-blown autonomy, the goal is to robustly obey an individual, regardless of the person’s pose, time of day, etc. Similarly, complex user experiences, such as gesture or emotional recognition, are not important for robotics that follow a human. In this sense, human-robot interaction reduces to simple "drone behavior programming" commands, such as determining whether to capture video from the person’s front or back, the angle and distance that follows the person, and simple "virtual fence" commands that prevent drone access to restricted areas. This program would be possible by manipulation of simple visual patterns.

Through this paper we want to put forward a method we developed to extend the use cases of drones/ UAV’s to sport activities, especially during running races using transfer learning which is a quick way to deploy a drone into the action. We provide a way to use a lightweight model to achieve the best trade off between performance and speed for object detection and applied computer vision techniques to mark the leading athlete and navigate the UAV according to this athlete’s movements.

In the next section we discuss about the related work done to solve similar tasks using various methods. Section 3 explains the approach we have implemented to solve this task. And section 4 explains the experimental setup which we have used to carry out our experiments and approach. The results of our work are discussed in section 5. Finally, we make a conclusion about our work and its extension in section 6.

# Literature Review

For this work we have referred many researches works of other people which share a similar aim and methodology, all of which are mentioned in the acknowledgement section of this paper. Primarily there are a few papers that tackle a similar problem using various computer vision techniques while others have used machine learning and deep learning methods. These methods include the usage of HOG feature vectors to identify the human like structures in an image and implement the k-means clustering algorithm, together which detected humans and their direction of motion in the frame. This is critical for reidentification of individuals which is another active research problem.

Another method using the computer vision techniques is usage of skpexels - a spatio-temporal depiction for skeleton sequences to fully utilize the "local" correlations between joints using the 2D convolution filters of Convolution Neural Networks. They converted skeleton videos into Skepxel-based images of flexible dimensions and develop a CNN-based structure for efficient human action recognition using the skpexel images.

Due to its robustness and high accuracy the method of deep learning was extensively used in object detection. For the task of human recognition, the current state of the art is Retina Net. Retina Net provides the greatest accuracy of human detection among all the deep learning approaches (Lin, Goyal, Girshick, He, & Piotr Dollar, 2018). In the paper, the images temporal relation has been used to enhance the human detection accuracy. Their task has been broken down into two sub tasks – firstly to detect if there are any humans in the image and next to identify their locations in the image. When a series of images is employed, the model accuracy of human detection has increased by 21.4 percent as compared to making use of only one image.

Another method known as Hierarchical Extreme Learning Machine (H-ELM), which is one of the unsupervised feature learning methods, uses sparse auto encoders to deliver more strong features that adapt to data variations without pre-processing. These deep neural models have proven to be skilled in human and non-human classification. Yet another research work approaches the task of face discovery using frames from the video and applied to the approval of the face detection, is a Haar-cascade classifier and max-margin object detection with CNN based features because they have high accuracy. To develop an obstacle detection system, Color discovery system has been used, which only focuses on the color of bodies and thereby detects the impediments in the way of drone.

In a different paper, the authors are exhibiting a self-governing drone having person detection and tracking framework which utilizes a static wide-edge camera and a lower-edge camera mounted on a pivoting turret. To utilize memory and time productively, they have proposed a joined multi-outline profound learning location procedure, where the casing coming from the zoomed camera on the turret is overlaid on the wide-edge static camera's casing. With this methodology, we can assemble an effective pipeline where the underlying discovery of little measured flying interlopers on the fundamental picture plane and their location on the zoomed picture plane is performed at the same time, limiting the expense of asset thorough recognition calculation. Using YOLO algorithm and CNN on NVIDIA GPU to train a deep learning model to detect humans/ persons or other objects on the frame. Now programming the drone movements, to follow the object in whichever direction it moves. Their future work included the detection of possible collisions.

# Approach

The main aim of this work is to make a UAV that can identify the leading athlete and follow that person on the racetrack. This task can be broken down into two parts – firstly, identifying the leading athlete and getting the locations of that person. Secondly, we must program the drone in such a way that the drone always follows the leading athlete only.

Preparing a model to recognize humans form scratch without any training would take a great many training data and hours or long stretches of training time. To speed this up, we can utilize Transfer learning – a process where we use the weights of the model that has been trained on a large amount of data to perform a analogous task. And then fine tune the layers in the way we want our results. Many models are available which we can use which have been trained to distinguish a wide variety of objects in images. From these trained models we can use checkpoints of their training phase and then apply them to our own task of detection. Transfer learning helps not only in cutting down the time required for training a model, but one can also improve the model’s accuracy by training the layers further on different data.

We decided to use MobileNet as it has a lightweight architecture. It uses depth wise separable convolutions which essentially means it performs a single convolution on each color channel instead of blending all three and flattening it. This has the consequence of filtering the input channels. This architecture also needs very low maintenance hence it performs well with high speeds. Table 1. Describes the architecture of MobileNet. SSD layers are added to the last DSC layer to replicate the architecture we have used.

The weights of the MobileNet are pretrained on COCO dataset. For detecting the athlete in each frame of a live stream video obtained from the drone, we are using COCO SSD MobileNet\_V1, a MobileNet neural network trained on the COCO dataset. MobileNetis an efficient Convolutional Neural Network for Mobile Vision Applications, Howard et al, 2017.COCO is a large object discovery and segmentation dataset. That has over 66 thousand instances of humans at various poses and under diverse lighting conditions. COCO expands to common objects in context, like the name says the images are taken from everyday scenes.

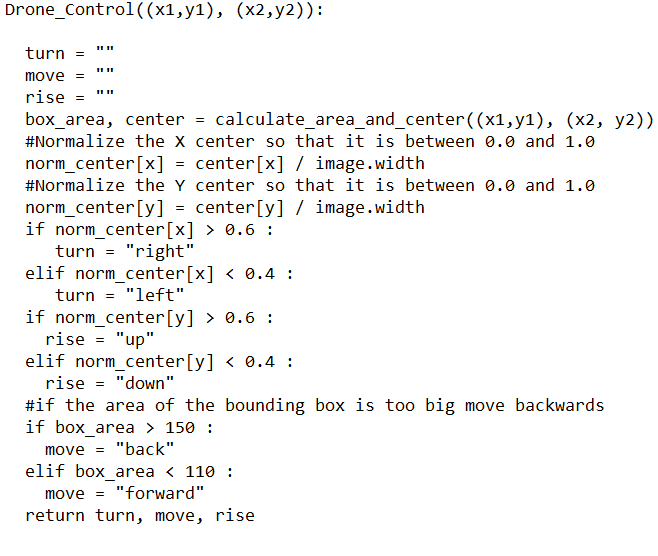
The authors of the MobileNet research paper demonstrated that ReLU6 is better than regular ReLU when we use low precision computation. Single shot detector (SSD) generates anchors and selects the topmost convolutional feature map and at a lower level it selects a feature map having higher resolution.

After that it adds a sequence of convolutional layers with spatial resolution with a specified configuration (decay rate for instance)

Using this architecture and pretrained weights on COCO we can obtain the locations of every person in the image and using coordinates we are able to mark a bounding box around each individual. Using these coordinates and computer vision techniques we can identify the individuals at the extreme ends of each frame. Depending on the direction of the race the drone will select either individual as the leading athlete.

Now, the second task of following the athlete throughout the race, begins. We tackle this task by measuring the relative position of the center of the bounding box with respect to the center of the frame of video stream. Fig. 2 We will move the drone/ UAV in such a way that these two points always overlap each other. Once the UAV has detected a leading athlete, it returns the 4 coordinates of the bounding box: its top-left corner symbolized by (X1, y1) and its bottom-right corner denoted by (X2, y2). Given these, we will compute the center of the box and its area. To compute the area, we compute the width as (X2-X1) and the height as (y2-y1) and multiply them. While the center, is calculated as (X2+X1)/2 and (y2+y1)/2

A pseudocode of the flight parameters to enable the UAV to follow the Athlete marked is shown below.



The total flow of the activities is as follows, first the UAV captures live video feed through its camera, then the video that is captured is broken down to frames which are processed one after the other. After which each processed frame is sent as an input to the SSD MobileNet model to detect the no. of persons. Then using the coordinates of each person we locate the athlete at the extreme ends of the frame and mark that person. This bounding box over the marked person will act a s an input to the drone navigation control which will command the drone to move in a way that the bounding box is always at the center of the frame.

The workflow can be described as shown in Fig. 1.

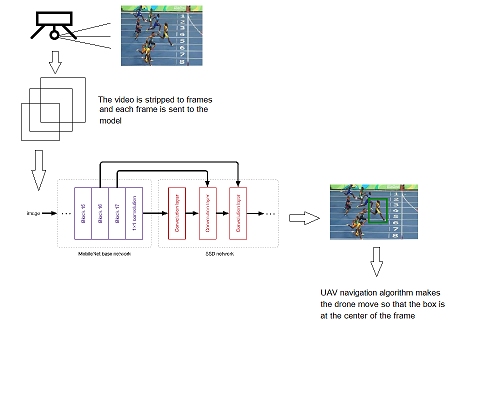


Fig. 1. Workflow of the approach which includes several steps. 1- The gimbal captures the video footage of the race from the side, 2- this video is broken down into frames and each frame is sent to the model as an input after which the leading athlete is marked with computer vision and finally 3- the UAV navigation algorithm commands the UAV to move accordingly.

# Experimental Setup

A person detection platform has been created using Python 3 and Tensorflow version 1.15.2 and recreated the model architecture to load the weights. The model has been downloaded from tensorflow model zoo which is publicly available. A pipeline of different processes at each stage was created. Initially the videos are captured and processed frame by frame and each frame has been resized before being converted to an array of pixels and sent to the model as input. After which the tensorflow graphs are loaded and the model makes a prediction. Then the model returns the coordinates of the bounding boxes (in the format [ymin. xmin. ymax, xmax]) which are used to calculate the position of each person identified in the frame. With these coordinates the leading athlete can be detected by calculating the position of each bounding box with respect to the center of the frame. The person at the extreme ends is marked as the leading athlete and a bounding box is drawn over the leading athlete only.

Coming to the hardware components i.e. the drone setup includes standard propellers – a pair of 1045 propellers to help pull the drone through air, with clockwise rotation. And pusher propellers at the back rotating counterclockwise. Brushless motors (Model: A2212 1000KV), which are better than a brushed motor, are used. Since we have a gimbal at the bottom it is mandatory to have ground clearance, for this purpose 4 pieces of tall landing skid gear legs are mounted. To control the speed of motors an electronic speed controller is used (30A ESC). The flight controller on the drone is a pixhawk 2.4.6 32bit ARM RC flight controller. A gimbal and gimbal controller board are used for capturing live feed through the drone and to send commands to the gimbal on the drone respectively. An RC receiver transmitter set (Fly Sky FS-i6 2.4G 6 channels AFHDS RC transmitter and a FS-iA6B receiver is used) to communicate with the drone and to transmit the video to the ground control. Collision avoidance sensor (TINY LIDAR Laser Ranging Sensor VL53L0X having 2 meters range), although not mandatory, helps from damaging the drone due to any obstacles during the flight. And a battery 7.4 V, 2200mAh is needed to power up the drone and keep all the components running. To make room for all these components, while making sure it is light weight, the frame of the drone (body) acts like a hub to place all these parts, so a frame model: F450 HJ450 DJI is used.





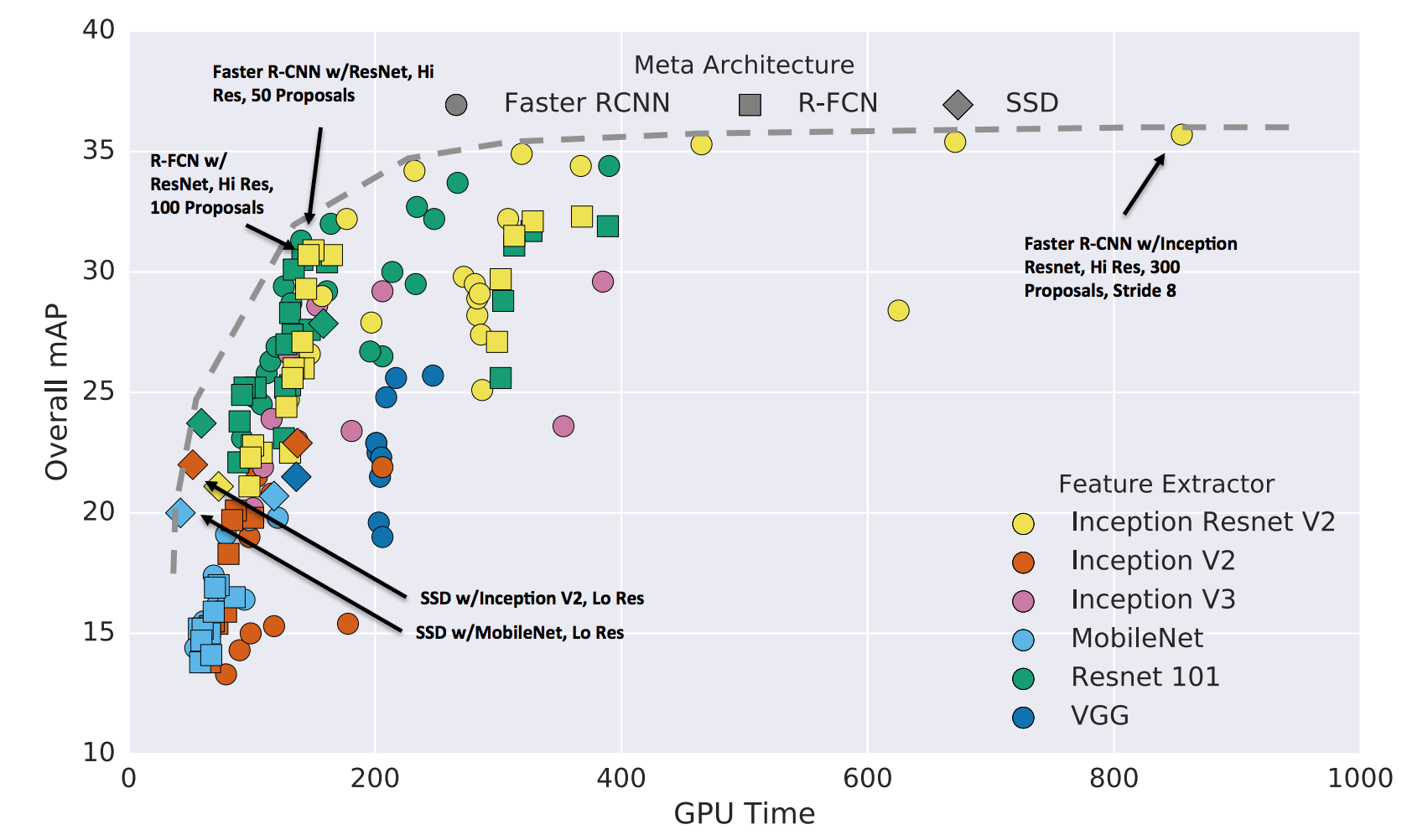


Fig. 2. In the above two instances the leading athlete is marked by a green bounding box while the trailing athlete is marked in red. The UAV will follow the green box in such a way that the center of the box is at the center of the frame. While the last image was used to detect the leading athlete when the race direction is configured.

# Results

As seen in Fig. 2 the model was able to detect most of the test cases. A few misclassifications happened due to the image quality and the form of athletes in the picture. units for each quantity in an equation. The classification accuracy among the test images taken from internet containing 100 random images of races from the side view came out to be 94.9 percent. The usage of SSD with Mobile-Net produced the best accuracy tradeoff with the performance speed. See Fig. 3 for the model accuracy when trained on COCO dataset over all the classes.

Single shot detector when used with MobileNet has achieved the highest mAP among all other advanced models when tested for real time processing, which explains its balance between speed and accuracy.



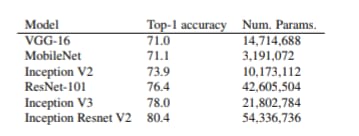




Fig. 3. The performance of the pretrained model on the coco dataset. The image on top shows the comparison between MobileNet and other popular architectures. Due to its light architecture (lesser parameters) the model performs very quickly compared to most state-of-the-art architectures while not compromising the accuracy. The comparison between the models with respect to GPU times is made by a Google research.

# Conclusion

In conclusion we would like to point out that usage of transfer learning for tasks like athlete following UAV, which needs to be quick at handling the frames and detecting humans, is very useful because it removes the training phase while deploying a drone into a required activity. Using transfer learning, in new domains having inadequate data or compute power for a model to train is, useful to implement many advanced neural networks that have been trained on a large amount data and have state of the art performance. We also want to point out that Mobile Net SSD trained on COCO is a fast and robust model for this task since it only has a footprint of around 1 GB in memory. Today most of the drones are being used for various tasks like aerial photography, security and surveillance and other military activities. This paper introduced a way to put UAV’s into work in sports domain. After this paper we would like to improve our current workflow to maximize the efficiency and reduce latency issues. In future we want to extend our work into other activities, especially in sports, where usage of drones may give us a better perspective in that domain.

References

1. A. J. A. Rivera, A. D. C. Villalobos, J. C. N. Monje, J. A. G. M. nas, and C. M. Oppus, “Post-disaster rescue facility: Human detection and geolocation using aerial drones,” in IEEE Region 10 Conference (TENCON), Nov 2016, pp. 384–386.
2. K. H. Ang, G. Chong, and Y. Li. PID control system analysis, design, and technology. IEEE Transactions on Control Systems Technology, 13(4):559–576, 2005.
3. J. Aulinas, Y. R. Petillot, J. Salvi, and X. Llado. The SLAM problem: a survey. In Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence, pages 363–371, 2008.
4. J. Aulinas, Y. R. Petillot, J. Salvi, and X. Llado. The SLAM problem: a survey. In Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence, pages 363–371, 2008.,   
   vol. 134, pp. A635–A646, Dec. 1965.
5. S. Avidan. Ensemble tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(2):261–271, 2007.
6. Person-following UAVs. <http://www.svcl.ucsd.edu/projects/dronefollow/>.
7. A. Borji and L. Itti. State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1):185–207, 2013.
8. Z. Cai, M. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 3361–3369, 2015.
9. P. Viola, M. J. Jones, and D. Snow, “Detecting pedestrians using patterns of motion and appearance,” in 9th IEEE International Conference on Computer Vision, vol. 2, Oct 2003, pp. 734–741.
10. R. Cowie. Emotion recognition in human-computer interaction. Signal Processing Magazine, pages 32 – 80, 2001.
11. H. Bristow and S. Lucey, “Why do linear SVMs trained on HOG features perform so well?” CoRR, vol. abs/1406.2419, pp. 1–8, 2014. [Online]. Available: http://arxiv.org/abs/1406.2419
12. M. G. Dissanayake, P. Newman, S. Clark, H. F. DurrantWhyte, and M. Csorba. A solution to the simultaneous localization and map building (SLAM) problem. IEEE Transactions on Robotics and Automation, 17(3):229–241, 20012.
13. D. E. King, “Max-margin object detection,” CoRR, vol. abs/1502.00046, pp. 1–8, 2015. [Online]. Available: http://arxiv.org/abs/1502.00046.
14. . M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascalnetwork.org/challenges/VOC/voc2012/workshop/index.html/
15. A. L. Yarbus. Eye movements and vision. New York: Plenum Pres, 1967.
16. W. G. Aguilar, V. P. Casaliglla, and J. L. Polit, “Obstacle avoidance based-visual navigation for micro aerial vehicles,” Electronics, vol. 6, pp. 1–23, 2017.
17. V. Mahadevan and N. Vasconcelos. Biologically inspired object tracking using center-surround saliency mechanisms. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(3):541–554, 2013.
18. P. Viola and M. J. Jones, “Robust real-time face detection,” International Journal of Computer Vision, vol. 57, no. 2, pp. 137–154, May 2004.
19. Bouabdallah, S. Design and Control of Quadrotors with Applications to Autonomous Flying. PhD thesis, Ecole Polytechnique Federale de Lausanne (2007).
20. Kumar, V. & Michael, N. Opportunities and challenges with autonomous micro aerial vehicles. Int. J. Robot. Res. 31, 1279–1291 (2012).
21. Shen, S., Mulgaonkar, Y., Michael, N. & Kumar, V. Vision-based state estimation and trajectory control towards aggressive flight with a quadrotor.
22. Brescianini, D., Hehn, M. & D'Andrea, R. Quadrocopter pole acrobatics. In Proc. International Conference on Intelligent Robots and Systems 3472–3479 (2013).
23. 2018, [online] Available: <https://www.marketsandmarkets.com/Market-Reports/video-surveillance-market-645.html>.
24. N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 886-893, 2005.
25. R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587, 2014.
26. R. Girshick, "Fast R-CNN", Proceedings of International Conference on Computer Vision, pp. 1440-1448, 2015.
27. P. Dollar, R. Appel, S. Belongie and P. Perona, "Fast Feature Pyramids for Object Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, pp. 1532-1545, 2014.
28. S. Pan and Q. Yang, "A Survey on Transfer Learning", IEEE Transactions on Knowledge & Data Engineering, vol. 22, pp. 1345-1359, 2010.
29. A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", Proceedings of International Conference on Neural Information Processing Systems, pp. 1097-1105, 2012.
30. Qichang Hu, Peng Wang, Chunhua Shen, Anton van den Hengel, Fatih Murat Porikli Computer Science IEEE Transactions on Circuits and Systems for Video Technology 2018.
31. Al-Kaff, A., Q. Meng, D. Martín, A. de la Escalera, and J. M. Armingol. 2016. “Monocular Vision-based Obstacle Detection/Avoidance for Unmanned Aerial Vehicles.” Intelligent Vehicles Symposium, IEEE, Gothenburg, Sweden, June 19–22.
32. Andert, F., and F. Adolf. 2009. “Online World Modeling and Path Planning for an Unmanned Helicopter.” *Autonomous Robots* 27 (3): 147–164.10.1007/s10514-009-9134-y.
33. Araar, O., and N. Aouf. 2014. “A New Hybrid Approach for the Visual Servoing of VTOL UAVs from Unknown Geometries.” Paper Presented at the Control and Automation, 22nd Mediterranean Conference of, Palermo, Italy, June 16–19.
34. Aulinas, J., Y. R. Petillot, J. Salvi, and X. Lladó. 2008. “The SLAM Problem: A Survey.” Conference on Artificial Intelligence Research and Development, 11th International Conference of the Catalan Association for Artificial Intelligence, Spain, October 22–24.
35. Bai, G., X. Xiang, H. Zhu, D. Yin, and L. Zhu. 2015. “Research on Obstacles Avoidance Technology for UAV Based on Improved PTAM Algorithm.” Paper Presented at the Progress in Informatics and Computing, IEEE International Conference on, Nanjing, China, December 18–20.
36. Bertrand, O. J., J. P. Lindemann, and M. Egelhaaf. 2015. “A Bio-Inspired Collision Avoidance Model Based on Spatial Information Derived from Motion Detectors Leads to Common Routes.” *PLoS Computational Biology* 11
37. Bonin-Font, F., A. Ortiz, and G. Oliver. 2008. “Visual Navigation for Mobile Robots: A Survey.” *Journal of Intelligent and Robotic Systems* 53 (3): 263–296.10.1007/s10846-008-9235-4.
38. Carrillo, L. R. G., A. E. D. López, R. Lozano, and C. Pégard. 2012. “Combining Stereo Vision and Inertial Navigation System for a Quad-Rotor UAV.” *Journal of Intelligent & Robotic Systems* 65 (1–4): 373–387.10.1007/s10846-011-9571-7.
39. Cho, D., P. Tsiotras, G. Zhang, and M. Holzinger. 2013. “Robust Feature Detection, Acquisition and Tracking for Relative Navigation in Space with a Known Target.” Paper Presented at the AIAA Guidance, Navigation, and Control Conference, Boston, MA, August 19–22.
40. Desouza, G. N., and A. C. Kak. 2002. “Vision for Mobile Robot Navigation: A Survey.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (2): 237–267.10.1109/34.982903.

1. [↑](#footnote-ref-1)