A Novel Aircraft Refueling Behavior Detection Model based on Deep Learning

Ran He1,*, Jianxing Wang2 and Zhengning Zhang3

1Aerospace ShenZhou Smart System Technology Co., Ltd., China Academy of Space Technology, Beijing, China
2China National Aviation Fuel Group Limited, Beijing, China
3Department of Electronic Engineering, Tsinghua University, Beijing, China

*Corresponding author e-mail: heran@spacesystech.com

Abstract. How to efficiently and accurately monitor the sparse aircraft Refueling behaviors from a large amount of video streams is of great significance for improving the level of management and refueling efficiency of aviation fuel stations. Due to the COVID-19 virus epidemic, the number of flights has dropped severely, the collection of image samples for refueling behaviors from large airport becomes difficult, which hinders the real-time detection of Refueling behaviors and reduces the efficiency of aviation fuel station. Therefore, automatically detecting the refueling behaviors of each station in time and accurately from a large number of aviation refuel stations still keeps challenging. To address this challenge, we propose a novel aircraft refueling behavior detection model based on deep learning, to quickly and accurately determine the refueling behaviors through analysing the video stream collected from the massive cameras deployed in the airport. Our proposed model adopts Inception v3 architecture of ImageNet to realize the model capability of transfer learning, the data augmentation to address the issue of over fitting, and the mAP (mean Average Precision) to test the performance. Our proposed model is also applied in the detection of refueling behaviors in China National Aviation Fuel Group, LTD (CNAF). The practical application results show better performance than other existing methods. Our work will promote the updating of related industry standard.

1. Introduction

Network cameras are widely used for some large enterprises or organizations in supervision and administration of production safety. For instance, CNAF has deployed a large number of network cameras in all the branches across China to form a complete video surveillance network that enables the personnel for the administration of production safety accessing working area visual data instantly [1]. Comparing to traditional supervision method, network cameras and vision network have the following three aspects of advances. First, it provides efficiency by allowing supervisors to access production information at the offices or anywhere having access to the vision network [2]. Second, through this kind of surveillance network, we can check the work situation at any time to ensure manufacturing operations comply safety specification. At last, the labor costs can be efficiently reduced, and unnecessary personnel contact can then be avoided, especially in the case of COVID-19 [3].
In spite of the advancement of the network cameras for CNAF, in order to quickly and accurately determine the refueling activities of each station, how to extract the useful information or features from the massive video stream data samples from the vision networks is a key challenge [4]. that is, the most important information from numerous video stream should be mined, instead of randomly selecting some steam or traversing all data [5].

Airplane refueling scenario is one of the most significant scenario in CNAFs vision network, however, there has hundreds of network cameras in an airport while seldom cameras are shotting at refueling scenario [6] [7] [8]. As it happens, it is urgent to find a method to pick out so called hot-spot cameras efficiently to support users to focus on these cameras [9]. The methods include two categories. One is to manually pick important cameras [10], the other is to adopt traditional machine learning methods, like Support Vector Machine (SVM), LR (Linear Regression), etc. [11], or Convolutional Neural Network (CNN) based deep learning (DL) methods to perform the classification tasks.

Apparently, it is inefficient to adopt manual method. While for machine learning or deep learning methods, the common process of image classification involves two stages: feature extraction and classification [12] [13]. For feature extraction, the differences between different categories should be captured [14] [15]. After that, a classification model will be trained by using these features and their corresponding class labels, the commonly used classification models include SVM, LR, RF (Random Forest) and Decision Tree [16], etc.

One of the main difficulties existed in this kind of traditional models is that, parameters during the process of feature extraction can not be adaptively adjusted when the images and their labels are changed [17]. If the features selected are insufficient to distinguish different categories, the accuracy of the model will be greatly reduced. The common methods for solving the problem are to use a variety of feature extractors, and then combine them to get a better performance, but a large amount of heuristic rules and manpower are then required to adjust parameters according to different fields to achieve a good accuracy, which is close to the human level. This is why it takes years to build a good computer vision system using traditional computer vision technology.

DL has been successfully applied in various domains, such as computer vision, speech recognition and natural language processing [18] [19]. CNN represents one of the most significant advances in DL due to its success in many challenging classification tasks [20]. CNN is a feed forward neural network, usually including feature extraction layer and feature mapping layer, and can learn local patterns in data by convolution. A distinctive property of CNN is that it is suitable for end-to-end learning without any a priori feature selection [21] [22]. Hence, lots of DL based classification researches and applications have emerged in recent years. However, vehicle refueling behavior detection model related research has not yet been reported [23].

In this work, we propose an improved CNN-based model to identify the vehicle refueling behavior by using the image data from CNAF vision network as training data and verify the performance of our proposed model to guarantee the airport production safety supervision.

2. Method

2.1 Backbone Selection

In this work, we mainly address the issue of quick and accurate recognition of refueling behaviors in each station in the case of the sparse distribution of fuel station location, thereby improving the utilization efficiency of aviation fuel stations. We mainly take the following aspects into account to improve our model. First, the maturity of the model should be considered in advance. Considering that the algorithm needs to be applied in the practical system, a mature algorithm model should be preferred rather than newly proposed network model. Second, regarding the relatively small of the number of data samples, the scale of model should be moderate. Too large of model scale would require a large number of data samples. In addition, the larger the model scale, the model computational performance will decrease. Third, to facilitate the improvement of model generalization, we design the model architecture with transfer learning to achieve better classification performance.

At present, the visual classification task mainly uses Convolutional Neural Networks (CNN). The early proposals of LeNet5 and AlexNet established the basic design pattern, which begin with
convolution layers, then pooling layers, and finally full connection layers. Then, the Visual Geometry Group from Oxford University proposed VGGNet [24] with the concept of using a standard 3*3 convolution kernel to replace the 7*7 convolution kernel, reducing the overall network scale with similar performance. Furthermore, The GoogleNet [25] [26] team proposed the Inception network structure, using multiple scales of convolution kernel combinations in the same layer, which is more close to biological visual processing. Accordingly, ResNets [27] [28] architecture proposed by He Kaiming solves the problem of gradient explosion caused by extremely deep network. Currently, most CNN based architectures designed for classification are optimized or improved based on the design ideas of the above.

This subject is mainly to solve the problem of binary classification network training in a small sample space. A deep network is not helpful to the problem. Therefore, the Inception series network is more suitable than the ResNets [29]. Compared to Inception V1 and V2, Inception V3 adopts the method of decomposing the convolution kernel to reduce the scale of parameters. At the same time, it adopts standardized super parameters, which greatly reduces the workload of network design. This is very suitable for quickly training an effective network model based on small samples. Inception V4 mainly optimizes the super parameters of Inception V3 and increases the depth of the network, which scale is larger than V3. In summary, the relatively small-scale Inception V3 is more suitable for the scenarios discussed in this article.

2.2 Construction of Inception V3 based Model

The architecture of Inception V3 is illustrated in Figure 1. It started with 5 convolutional layers and 2 max pool layers, then followed with 11 Inception structures. Finally, output the result with 1000 parameters through full-connection layers and softmax function. The size of the Inception of V3 is relatively huge. But we only collected 671 positive sample pictures. There is a big gap between the amount of sample data and the network scale. Therefore, a pre-trained model on the open-source data set should be considered first, we could adopt the method of transfer learning to reduce the number of parameters that need to be trained. We choose the Kaggle based Inception V3 pre-training model.

![Figure 1. The architecture of Inception V3.](image)

The Inception V3 network, which original solving 1000 classification problems, contains 10 Inception processing procedures, then followed by the pooling layer and the softmax layer. In this paper, the model needs reforming to a binary output one. Considering that the Inception V3 network has extracted features through multi-layer convolution, there is no need to add convolution and pooling layers. Thus the model structure should update like follows: First choosing an appropriate output layer in Inception V3, then reduce the output to a 1-dimensional vector, and use the fully connected layer to reduce the number of output parameters, and use ReLU to activate the output. Finally, fully connect the activated output parameters to 1 parameter, and put it through Sigmoid function. The construction of the optimized transfer learning model is illustrated in Figure 2.
Flatten. The input data is retrieved from the mix7 Inception structure of Inception V3. While input images size is 352*288, the input data size is 20*16*768, which is suitable for this research. The flatten layer converts the output to a 1-dimensional vector with 245,760 parameters.

FullConn. The fully connected layer aims to perform high-level reasoning on feature representation. FullConn takes all neurons in Flatten and connects them to 512 neurons of the current layer to generate a global semantics input picture. The dropout strategy is applied to prevent overfitting. The output of each hidden neuron in FullConn is set to 0 with probability 0.4. The dropout strategy forces the model to learn more robust features.

Output. Logistic regression is put on top of the previously hidden layers as the output layer of the classifier. We used sigmoid as the activate function to transfer the result to a probability value between 0 to 1.

Figure 2. The improved architecture of our proposed model based on Inception V3.

2.3 Dataset
The data set used to train the AI model will directly affect the results of the algorithm and play a vital role in the ultimate users experience. A training set which has a similar distribution with the real world scenario can improve the accuracy of the algorithm.

To improve the efficiency of data collection, the research team wrote a script automatically collect training image data by periodically visit more than 200 cameras at an airport. The strategy can also help to make data set more related to a real-world scenario. However, more effort should be put into consideration by analyzing the actual production scenario. The specific points to be considered are as follows:

1) Avoid similar data. Due to the fixed installation of the apron camera, the shooting angle rarely changes. For the same camera, in most cases, the picture almost remains the same. Therefore, the time intervals between collecting images from the same camera should not be too short, about 1 hour is appropriate;

2) Consider the changes in lighting conditions caused by weather and time. For the same camera, weather conditions, and the angle of sunlight in the dawn and dark will have a certain impact on the image. The collected images need to include different weather, time, and sunshine conditions.

3) The busyness of the airport stand. The camera for the long-distance airport stand may have no flights arriving for a long time, and the corresponding refueling scene images are less or missing, which leads to the discrimination of such scenes by the algorithm. Although the deep learning algorithm has good generalization ability, it is still necessary to avoid such problems in the data collection phase, and the amount of data from some busy camera needs to be carefully reduced.

4) Unbalance between positive images and negative images. Commonly, the data obtained by the automatic polling script is a negative sample of “no refueling operation”. There only has 671 positive samples in the 9821 images we collected. This would cause the algorithm to tend to choose
negative without interfere. One way to solve this problem is to simply reduce the number of negative samples. We can also choose to adjust the design of the training loss function to increase the penalty for missed detection of positive samples.

Considering the constraints of the training data and we have only about 200 cameras available for collecting data, the number of samples will be greatly restricted. It is worth mentioning that because of the outbreak of the COVID-19, the number of flights decreased, thus the refueling operation scene was rare. This brings greater difficulties for data collection. After two months, only 671 valid positive sample images were finally collected while we have 4951 negative samples.

Considering the serious imbalance between the positive sample and the sub-sample, if all the negative samples are simply included in the training set, our problem will be a typical asymmetry classification problem. But, due to the character-istics of apron operations, a large number of the negative samples have great similarities (While there is no refueling operation, the camera usually shoots an empty apron or a stationary aircraft parked in a fixed position). It would lead to an increasing calculation burden, while has little effect on the variety of sample data. Therefore, we randomly disposed of negative samples until only 671 pictures left, which is the same size as positive samples. This effort can improve the training speed, and not cause too much negative effect on the training accuracy.

We manually labeled the samples, and split positive and negative data separately into the training dataset including 571 pictures, and validating dataset including 100 pictures.

2.4 Training

Due to the small amount of data, and transfer learning could radically reduce the amount of computing, we can use mid-end or low-end workstations for training. The training workstation is equipped with dual GeForce GTX TITAN X GPU, Intel Core i7-5930K with 6 physical core, and 64G RAM.

**Basic super parameters.** Training the classifier can be regarded as solving a typical optimization problem. Applying combined strategies to the training phase is necessary. We describe some practical strategies which we used during training in this subsection. Image size. Inception V3 recommends that the input image resolution is above 299*299. The size of our original image is 352*288. Our customized model takes the output from the middle Inception V3 mix7 layer, which reduces the image size requirements. As a result, the original image can meet the requirements. Learning rate. The BP (back propagation) algorithm provides an approximation of the trajectory calculated by using the steepest descent in the weight parameter space. The learning rate is initialized to 0.001 at the beginning. The batch size needs to be selected in combination with the learning rate and the performance of the training device, which will affect the speed of convergence. In our subject, the batch size is set to 30, which means 43 batches per epoch.

**Data augmentation.** The purpose of data augmentation is to increase the amount of data, enrich the data diversity, and improve the generalization ability of the model. The general approach of data augmentation is to alter the original data through operations such as rotation, width/height shift, shear, zoom, and flipping. It can alleviate the problem of over-fitting caused by the relatively insufficient data set.

The effect of image augmentation also depends on the parameters we chose. If the result of the image amplification is more consistent with the actual scene, a better effect can be exerted. In our scenario, the camera will adjust the angle or change the focal length to enlarge the view. The position of the aircraft and fuel truck usually shifts slightly. These are the fact we should consider. We describe some parameters which we used during data augmentation.

**Rotation range.** The choice of the rotation range depends on the images in the actual scenario. Vehicles and airplanes are placed horizontally, and the camera will shake slightly due to wind speed, so the actual rotation angle will not be too much. As a result, the rotation range is set to 5, which indicates that the degree range of random rotation is from -5 to 5.

**Width/Height shift range.** Width/Height shift makes images shift in a vertical or horizontal direction. Workers usually put the fuel vehicle in the middle of the monitoring scene, so the range of
horizontal and vertical shifts should not be too much. We set it to 0.1, that is, the range of horizontal or vertical movement won’t exceed 10% of the image size.

**Zoom range.** Apron cameras usually zoom in to see the detail of the operation or zoom out to get a wide range. Zoom range indicates the lower size is 1-zoom range, and the upper size is 1+zoom range. In this method, the zoom range is set to 0.4, which is a relatively large value, to simulate cameras zoom operation.

3. **Results and discussion**

Here we evaluate the predictive performance of the Inception V3 base model on both loss function and mAP. We use raw data and data amplification for training respectively, to compare the difference that the data augmentation could make.

Binary cross-entropy was chosen as the lost function, which is defined as:

\[
H(p, q) = \sum_x p(x) \log \frac{1}{q(x)} = -\sum_x p(x) \log(q(x))
\]

Where \(p(x)\) is the probability when the sample is \(x\), \(q(x)\) is the probability when the model predicts sample as \(x\). The value range of \(x\) is \([0,1]\), 1 indicate the picture is a positive sample, which means that the refueling operation. 0 indicate the picture is a negative sample, which means there is no refueling operation. When \(q(x)\) equals \(p(x)\), \(H(p,q)\) reaches its minimum value.

Figure 3 shows the resulting training with raw data. The training of the proposed model quickly converges to an mAP of 1.00, but the result on the test set is only 0.83, which is over-fitting. In Figure 4, it can be seen that by applying image augmentation, the training convergence speed is significantly reduced, and finally convergence around 0.96 accuracies. The mAP in the test set reaches 0.93, The over-fitting problem is solved, and the verification results meet the needs of the production system.

![Figure 3. The training results with raw image data.](image)

![Figure 4. The convergency of our proposed model by applying image augmentation.](image)

We deploy the model in a Docker container and integrate it with the Production Safety Supervision System(PSS). The PSS takes screenshots of scenes from each camera at a certain time interval, pushes the picture to the algorithm, and the algorithm returns the classification results. In the human-computer interaction interface, the PSS uses high light color to mark the cameras that are classified as refueling operations (Figure 5). System users can quickly find the cameras they focus on among hundreds of cameras. As far as we know, there is currently no related research for the recognition of refueling vehicles. By studying the vehicle refueling behavior detection model based on a small sample data set,
we proposed the first domestic aircraft refueling vehicle recognition algorithm in practice and improved the efficiency of safety product supervision in the airport.

Figure 5. The practical deployment application of PSS in CNAF based on our proposed model.

4. Conclusion
In our case, by analyzing the characteristics and principles of mainstream visual classification models, Inception V3 was finally selected as the backbone for training. Then we used the transform learning method. By fine-tuning appropriate parameters, 93% mAP was reached, which met the needs of practical applications. The model training method given in this article is of great reference value for the recognition of other proprietary scenes, such as heavy rain, oil spill, etc. It helps solve this kind of classification problem with a relatively small specific data set, and finishing training and optimization DL model in a short time.

5. Acknowledgments
We would like to thank CNAF for providing analysis data support.

References
[1] H. Wang, X. Dong, J. Xue, J. Liu, and J. Wang, Modeling and simulation of a time-varying inertia aircraft in aerial refueling, Chinese Journal of Aeronautics, vol. 29, no. 2, pp. 335 – 345, 2016.
[2] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, Dota: A large-scale dataset for object detection in aerial images, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3974 – 3983.
[3] M. Loey, F. Smarandache, and N. E. M. Khalifa, Within the lack of chest covid-19 x-ray dataset: A novel detection model based on gan and deep transfer learning, 2020.
[4] Y. Lu, System and method of establishing a multi-camera image using pixel remapping, Feb. 20 2018, US Patent 9,900,522.
[5] Y. Jiang, W. Wang, and C. Zhao, A machine vision-based realtime anomaly detection method for industrial products using deep learning, in 2019 Chinese Automation Congress (CAC). IEEE, 2019, pp. 4842 – 4847.
[6] E. Plaza and M. Santos, Knowledge based approach to ground refueling optimization of commercial airplanes, Expert Systems, p. e12631, 2020.

[7] W. L. Dunn Jr, C. J. Unrau, V. L. Tindell, and T. W. Hanke, Aircraft refueling safety system, May 17 2016, uS Patent 9,340,298.

[8] A. T. W. Min, R. Sagarna, A. Gupta, Y.-S. Ong, and C. K. Goh, Knowledge transfer through machine learning in aircraft design, IEEE Computational Intelligence Magazine, vol. 12, no. 4, pp. 48 – 60, 2017.

[9] R. Mash, B. Borghetti, and J. Pecarina, Improved aircraft recognition for aerial refueling through data augmentation in convolutional neural networks, in International symposium on visual computing. Springer, 2016, pp. 113 – 122.

[10] Z. Chen, T. Zhang, and C. Ouyang, End-to-end airplane detection using transfer learning in remote sensing images, Remote Sensing, vol. 10, no. 1, p. 139, 2018.

[11] J. Ahn, J. Park, D. Park, J. Paek, and J. Ko, Convolutional neural network-based classification system design with compressed wireless sensor network images, PloS one, vol. 13, no. 5, p. e0196251, 2018.

[12] Y. Wang, X. Chen, H. Cai, and Z. Liang, Effect of various image information in polyp classification by deep learning with small dataset, in 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP). IEEE, 2019, pp. 1 – 4.

[13] N. Huang, J. He, N. Zhu, X. Xuan, G. Liu, and C. Chang, Identification of the source camera of images based on convolutional neural network, Digital Investigation, vol. 26, pp. 72 – 80, 2018.

[14] D. T. Nguyen, T. D. Pham, N. R. Baek, and K. R. Park, Combining deep and handcrafted image features for presentation attack detection in face recognition systems using visible-light camera sensors, Sensors, vol. 18, no. 3, p. 699, 2018.

[15] Y. P. Wang, T. Xian, M. V. Miraglia, T. Williams, D. Ellis, J. Nahill, Y. Liu, P. A. Giordano, and D. Wilz, Method of using camera sensor interface to transfer multiple channels of scan data using an image format, Jun. 28 2016, uS Patent 9,378,403.

[16] A. Jalal and S. Kamal, Improved behavior monitoring and classification using cues parameters extraction from camera array images. International Journal of Interactive Multimedia & Artificial Intelligence, vol. 5, no. 5, 2019.

[17] B. Liu, X. Yu, P. Zhang, A. Yu, Q. Fu, and X. Wei, Supervised deep feature extraction for hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 4, pp. 1909 – 1921, 2017.

[18] A. Maier, C. Syben, T. Lasser, and C. Riess, A gentle introduction to deep learning in medical image processing, Zeitschrift f¨ur Medizinische Physik, vol. 29, no. 2, pp. 86 – 101, 2019.

[19] H. Ping and S. Qin, A social bots detection model based on deep learning algorithm, in 2018 IEEE 18th International Conference on Communication Technology (ICICT). IEEE, 2018, pp. 1435 – 1439.

[20] R. Mash, N. Becherer, B. Woolley, and J. Pecarina, Toward aircraft recognition with convolutional neural networks, in 2016 IEEE National Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS). IEEE, 2016, pp. 225 – 232.

[21] S. Yu, Y. Wu, W. Li, Z. Song, and W. Zeng, A model for fine-grained vehicle classification based on deep learning, Neurocomputing, vol. 257, pp. 97 – 103, 2017.

[22] K. W. Kim, H. G. Hong, G. P. Nam, and K. R. Park, A study of deep cnn-based classification of open and closed eyes using a visible light camera sensor, Sensors, vol. 17, no. 7, p. 1534, 2017.