International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 2, Mar-Apr-2020, ISSN (Online): 2395-566X

Multi-Model Attribute Generator (MAG). A Deep Learning Method for Attribute Label Generation for Images

Matthew Charles Millar
Research and Development
Unifa
Scottsdale, USA
matthew.millar@unifa-e.com

Abstract – **This paper looks to create a system that will allow for accurate automatic generation of labels for image data of people designed explicitly for ReID problems. There is a lack of research on attribute data generation for images, especially in ReID datasets. Most attribute data are hand-generated and not created from a network. This approach is labor-intensive and prone to errors as well as costly and time-consuming. This paper utilizes a unique method of multiple models to generated separate portions of the attributes for each image. The combination of multiple models allows for a highly accurate method for labels for attributes to be generated. The MAG system has shown to be accurate for particular attribute generation. Gender, Upper-body clothing type, and Object detection are incredibly accurate. While, the other attribute each has outstanding accuracy, just not to the same level as the top three models.**

Keywords – **Computer Vision, Pattern Matching, Deep learning, attribute labeling.**

I. INTRODUCTION

Generating accurate data annotations take a very long time, especially if building out an extensive dataset for training a computer vision model from scratch. The total amount of data can be in the thousands of images for a particular task. When dealing with two or three classes, a few hundred pictures of each will be enough. However, when working with more than 100 classes, the data required exponentially increases. This data increase makes annotation time consuming and sometimes extremely difficult to manage alone ([1], [2]).

II. BACKGROUND

It is common to use low-level features for image retrieval to aid in semantic understanding of an image. Correct feature representation can improve semantic learning in computer vision tasks, especially in image retrieval and automated annotation [3]. Image segmentation is a useful method for feature extraction in image annotation. The segmentation algorithm can section out aspects of an image into parts and categorize these parts as items in the image [4].

These parts are useful for annotating the image piece by piece. Much like image segmentation, a clustering algorithm will find similar pixels of an image and group them using a k-means method [5]. The issue with this approach is the total number of clusters must be known before segmenting the image. Edge detection is another approach for feature extraction [6]. This approach has useful applications, but only to simple images and gives a poor performance on complex images.

Color is a common attribute for feature extraction. Color can be considered one if not an essential feature of an image ([3]). The extraction of color features usually results from defining a color space. Color histograms are possibly the most used method for color feature extraction ([5]). The color histogram will describe the distribution of color in an image ([7]). This will give a description of the whole or sections of an image to use in its annotations.

Textures of an image are another aspect that can be an annotation. Describing the texture of a material can be beneficial in telling different surfaces apart from one another based on a comparison between two images. Texture measurements are typically taken from multiple adjacent pixels depending on the size of the image or the complexity of the image ([8], [9]). Texture extraction can fall into two main categories spatial texture features and spectral texture features. Spatial textures look at either pixel statistics or a local pixel cluster ([10]). Spectral extraction involves extracting features from the color frequency of an image ([11]).

The location of one object to another is also another critical aspect of image annotation and image retrieval. Spatial descriptions for annotations can be specific as to where the object location on the image. This spot is the absolute location of the object in an image ([12]). The description can also be relative (left or right). This approach is based on ontology and can be useful, especially if the object is limited in potential positions

([13]). Alternatively, even simpler, it can be a simple present or not a present (binary) descriptor.

The use of CNN and other deep learning models have been gaining traction in different fields of generative attributes. The most common application would be for facial features ([14], [15]). They have also been useful in image captioning for scene analysis or image description [16]. Furthermore, even for learning attributes to aid in Re-Identification projects [17]. However, these papers use attributes to build their models and dataset, not use a model to create the attributes for labels. This study looks at filling this void by creating a deep learning model that can accurately create attributes for a dataset of human images.

III. DATA

The dataset that was used in this study was collected from the standard Re-Identification datasets, Market1501 [18], and CUHK [19]. The primary dataset will be the Market1501, as this is one of the more common and readily used datasets in ReID studies and programs [20]. The CUHK dataset is used as a supplementary dataset for aspects that the Market1501 dataset did not have. The main issue was with long versus short sleeves. There were not enough long-sleeve images in the Market1501 to train a deep learning model. Due to the lake of long sleeve shirts, this study use of CUHK dataset to supplement the long-sleeved shirts or jackets.

IV. METHODS

The methodology that is used in the paper is to break down the problem and solve it using multiple models to generate the attributes. This method allows for several smaller models to be built and trained with less data than the use of a deep end to end model. This methodology also allows for each model to be more accurate as it is not contending with other labels or other extractor and combination of layers, which can make a model overly complicated and not as successful for these more straightforward problems. The first step is to break down which parts will be used for attributes and then which attributes should be extracted. The main attributes that are considered are; the gender of the person, what is worn on the upper body, what is worn on the lower body, upper body color, lower body color, and if they have a bag or not. These attributes where all created into their own smaller models using different and appropriate techniques.

V. MODELS

This section will cover the structure, training, and setup of each model that was built to be used in the MAG system.

1. Gender Model

Gender is the first model that will be discussed. At first, this seems very simple as it is purely a binary classification model (male or female). Gender models have been primarily for facial recognition. Gender models have not been done for the whole body, let alone a slightly blurry image of a person at a distance [21]. With a finetuned Resnet50 model as the base extractor, a few Fully Connected Layers was built on top of the extractor, which would then produce a binary output (male/female) as well as fine-tune the feature extraction.

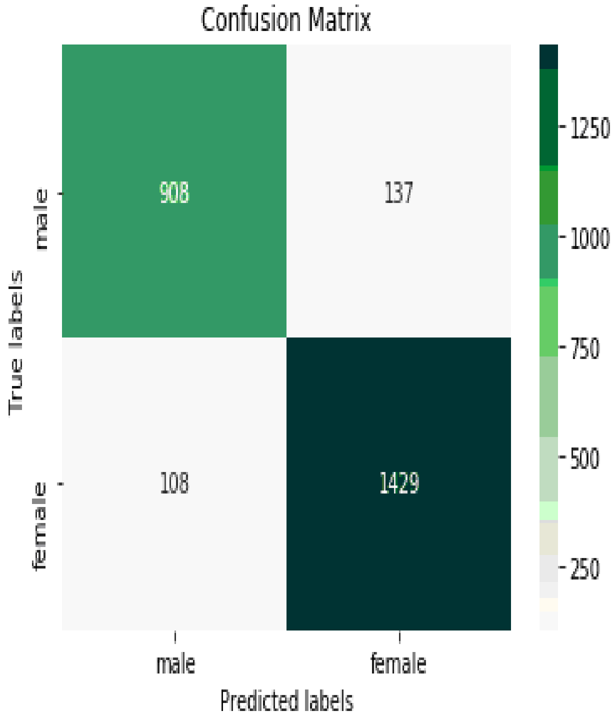

Fig.1. Gender Model Confusion Matrix

Figure 1 shows that the model was very successful at predicting the gender of a person with only a total of 245 misclassified images out of a possible 2337 giving an overall accuracy of 90% for gender identification.

2. Upper-Body Model

The second model will be the Upper-body Model. This model's purpose is to define what type of clothing a person is wearing on their upper body. The type of clothing can either be a shirt or tee-shirt (long or short sleeves). This approach is a bit genialized, but due to the quality of the images, the difference between types of shirts became very difficult to distinguish and prone to errors. Hence, the separation between only long sleeves and short sleeves, shirts/dresses. However, the dataset that was being used had a minimal collection of long sleeves shirts so that the dataset was augmented to include another commonly used ReID dataset, CUHK [22]. This new dataset gave the missing data form the Market1501 dataset. Much like gender descriptor, this model is another binary output. The upper-body model shares a similar model to the gender model, with different inputs. Looking at this model only wants to know the upper body; the input image is split in half only to include the upper portions of each image. This process was completed using a custom image cropping script.

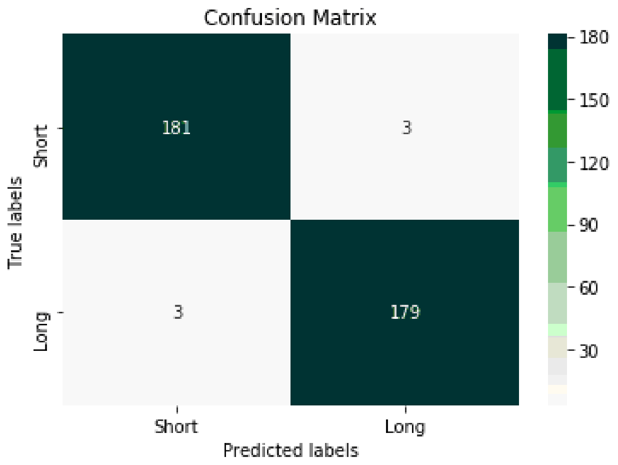

Fig.2. Upper-Body Model Confusion Matrix.

Looking at figure2, the accuracy of this model is very higher. There were only six misclassified images, resulting in a 98% accuracy overall.

3. Upper-body Color Model

The color of the upper body was based on a CNN, which could decern the primary color of the target. Instead of using the finetuned Resnet50 model of the previous models, this CNN was trained from scratch based on the architecture from Zhang research in vehicle color recognition [23]. With using a similar approach, excellent results were found. With a final accuracy of 79%, the model could infer if the primary color of a person was one of seven significant colors (determined from the Market1501 dataset). These colors were; red, white, yellow, green, purple, blue, and black.

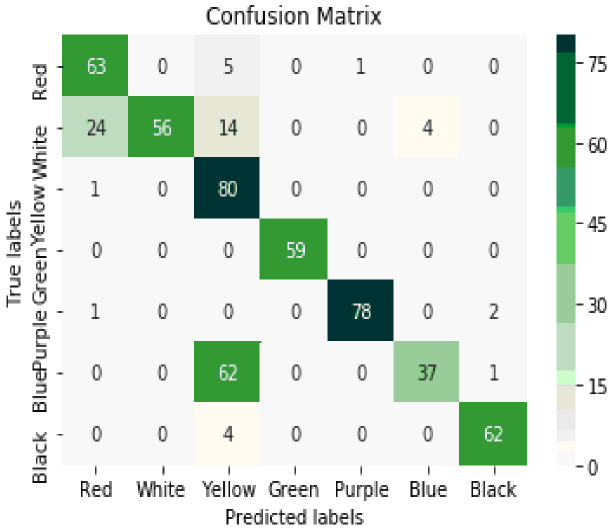

Fig.3. Upper-Body Color Model Confusion Matrix.

Some colors performed much better than others. Green, for instance, was 100% accurate. However, white and blue were the weakest performers in this model. Blue was mainly predicted as yellow, while white was mainly misclassified as red. This lower accuracy was due to the 'in-between' colors like teal, and pink, which proved to be difficult for the model to decern between these hard classes.

4. Lower-body Model

The lower body model was looking at three types of clothing. The model considered shorts, skirts/dresses, and pants. Much like the upper body model, the images are cropped down to just the lower portion of the body. This model was built on a finetuned resnet50 model that was trained on the Market1501 dataset. Then two blocks of FC layers were attached on top of the global average pooling layer. The use of batch normalization and dropout layers help manger the overfitting of the model due to some of the short and shirts being very similar. There was also an issue with some shorts being more like pants due to the length of the shorts. This slight difference between men and women cloting caused most of the misclassifications in the training of the model.

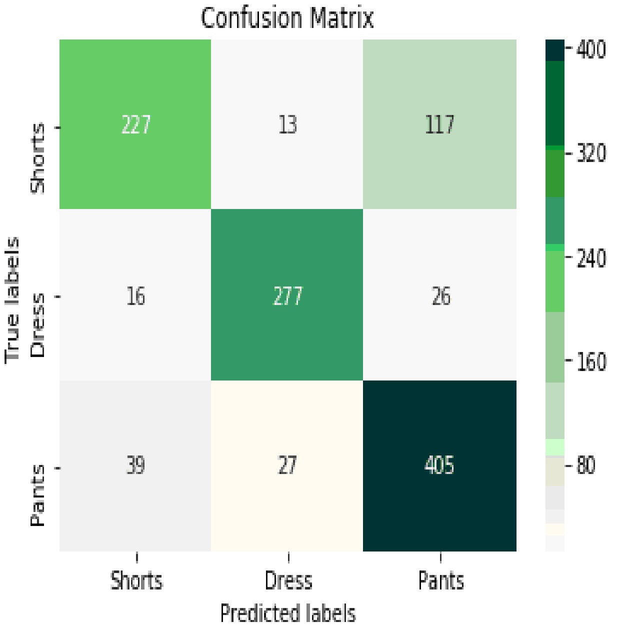

Fig.4. Lower-Body Model Confusion Matrix.

Figure 4 shows that the shorts class was mainly misclassified as pants which led to further investigation to show that the length of the shorts, and some pants were close. In some cases, women whose pants were rolled up looked very close to some men's shorts that were much longer. With an overall accuracy of 79%, this model performed well enough but still has some room for improvements.

5. Lower-body Color Model

The lower body-color model was the same CNN that was used in the upper body model but retrained for lower body colors, which were slightly different from the upper body.

The colors that were considered for the lower body were grey, white, green, black, red, blue, brown, and yellow. This model used a similar preprocessing method as the lower body model, the model was trained, giving an accuracy of 61%. This model was the lowest scoring of all the models in the MAG system.

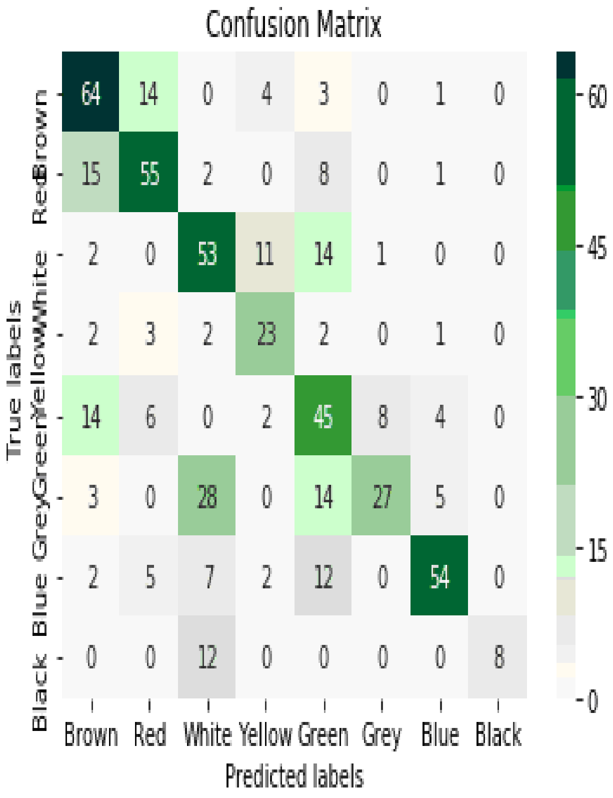

Fig.5. Lower-Body Model Confusion Matrix.

These lower scores may be due to the excess of background noise in the images compared to the upper body.

6. Object Detection Bags

The last aspect that was looked at was the detection of if there was a bag or not. This would include backpacks, handbags, or other carrying objects. This was done by retraining a Yolov3 [24] (Redmon & Farhadi, 2018). This approach was chosen as object detection is very well defined, and YOLOv3 is quick as well as accurate. YOLOv3 was chosen as it already has been trained on backpacks, handbags, and suitcases, which would cover the majority if not all the cases that are present in the Market1501 data. Using the pre-trained weights for YOLOv3 yields very good results. There was a small misstep in that when a bag was the same or very similar to the color of the upper body, and there was a miss in the detection. However, the majority of images were correctly identified, giving 93.38% accuracy in finding and correctly identifying the object that a person is carrying.

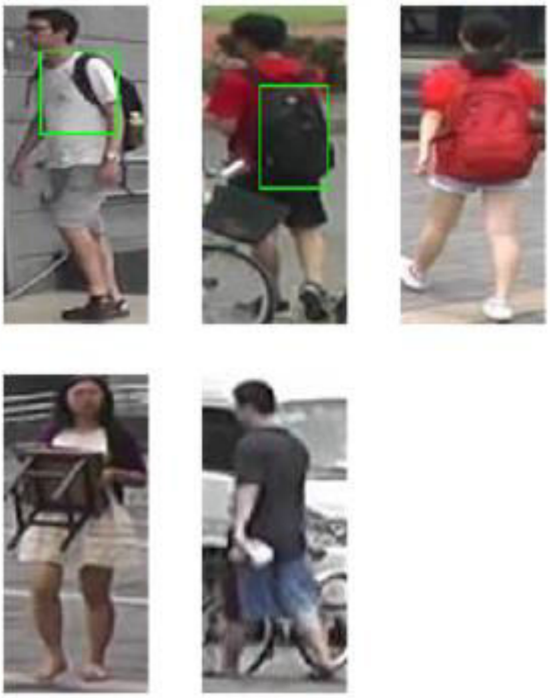

Fig.6. Bag Detection Model Visualization.

VI. RESULTS

The output can come in several forms depending on the project needs. For example, looking at this image:

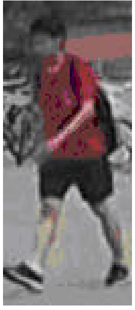

Fig.7. Example Image.

The resulting output was:
Sex: Male
UB_T: Short Sleeves
UB_C: Red
LB_T: Shorts
LB_C: Black
OBJECT: Backpack

Where sex is the gender, UB_T is the type of clothing on the upper body, UB_C is the upper body color, LB_T is the lower body type of clothing, LB_C is the color of the lower body, and OBJECT is the type of object if any they are carrying. The OBJECT would be limited to backpack, suitcase, and handbag, as that is what the YOLOv3 model would have detected.

The testing set consisted of randomly sampled images from both the Market1501 and CUHK datasets. A total of 400 images were selected for evaluation purposes. The results from the analysis were slightly better than the validation set analysis during the training phase.

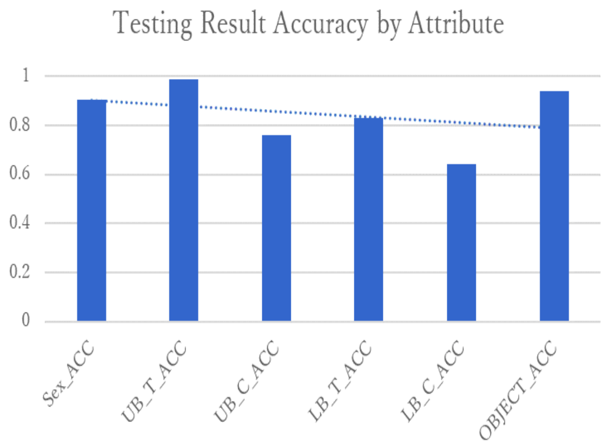

Fig.8. Accuracy by Attribute.

Looking at the results, the model's accuracy at providing labels for each randomly selected image is quite suitable for each label. The overall average label generation for each model is 84.5% accurate for all labels generated per image. The median and mode where both 83.33% with a standard deviation of 0.14. These results show that there are good grouping and accuracy for all images in the testing dataset.

VII. CONCLUSION

By combining all these model's outputs into a single output, an excellent description of a person can be generated. This description will have; the person's gender, what type of clothing they are wearing, what the color of the clothing they are wearing, and if they are carrying a bag or not. This information can then be written out into a JSON, XML, or CSV file along with the ID of the file. This newly generated dataset can then be used for building labels and descriptors for individuals quickly and accurately. This can be done much quicker compared to the hand-labeling image, which is time consuming, costly, and can be prone to mislabel due to differences of options between labelers. This mislabeling, which may never be caught, can cause issues with DL and ML model's training. Incorrect labels can slightly skew the results of a model even if there a large dataset and data augmentations are used [25].

The architecture of MAG is one of its key components that improve its success. By using single-task models, each task is completed more accurately over an end to end model, which would be far more sophisticated while producing more inferior quality results. Another benefit of MAG is the on the fly addition or subtraction of specific descriptors when needed. One small drawback to the multi-model approach is the preprocessing step for each image. The gender model and the bag detection model take in a whole image. While the color models and top and bottom will have to have the images divided into two equal parts. By using an end to end systems, the management of the preprocessing step will be minimal. However, MAG's overall accuracy gained would be much higher than the use of the end to end system or a multi-label classification CNN.

REFERENCES

[1]. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, L. (2014). Microsoft COCO: Common Objects in Context. Computer Vision – ECCV 2014 Lecture Notes in Computer Science, 740-755.

[2]. Koller, O., Ney, H., & Bowden, R. (2016). Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3793-3802.

[3]. Zhang, D., Islam, M., & Lu, G. (2012). A review of automatic image annotation techniques. Pattern Recognition, 45, 346-362.

[4]. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. The European Conference on Computer Vision (ECCV), 801-818.

[5]. Wang, J., Li, J., & Wiederhold, G. (2001). Simplicity: semantics-sensitive integrated. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(9), 947-963.

[6]. Schindler, M., Wegner, J., Galliani, S., Datcu, M., & Stilla, U. (2018). Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS Journal of Photogrammetry and Remote Sensing, 135, 158-172.

[7]. Huang, X., Zhang, R., Jia, K., Wang, Z., & Nie, W. (2018). Taxi Detection Based on the Sliding Color Histogram Matching. 2018 IEEE 3rd International Conference on Image, Vision, and Computing (ICIVC), Chongqing, 86-90.

[8]. Liu, P., Guo, J., Chamnongthai, K., & Prasetyo, H. (2017). Fusion of color histogram and LBP-based features for texture image retrieval and classification. Information Sciences, 390, 95-111.

[9]. Qi, X., Li, C.-G., Zhoa, G., Hong, X., & Pietikainen, M. (2016). Dynamic texture and scene classification by transferring deep image features. Neurocomputing, 171, 1230-1241.

[10]. Gonzalez, R. C., & Woods, R. E. (2017). Digital Image Processing, Global Edition (4th ed.). New York, NY: Pearson Education Limited.

[11]. Mirzapour, F., & Ghassemian, H. (2015). Improving hyperspectral image classification by combining spectral, texture, and shape features.

International Journal of Remote Sensing, 36(4), 1070-1096.

[12]. Fan, J., Gao, Y., Luo, H., & Xu, G. (2004). Automatic image annotation by using concept-sensitive salient objects for image content representation., (pp. 61-368).

[13]. Mazaris, V., Kompatsiaris, I., & Strintzis, M. G. (2003). An ontology approach to object-based image retrieval. Proceedings 2003 International Conference on Image Processing, (pp. 511-514).

[14]. Zhong, Y., Sullivan, J., & Li, H. (2016). Face attribute prediction using off-the-shelf CNN features. 2016 International Conference on Biometrics (ICB).

[15]. Li, M., Zuo, W., & Zhang, D. (2018). Deep Identity-aware Transfer of Facial Attributes. arXiv:1610.05586.

[16]. Wang, Y., Lin, Z., Shen, X., Cohen, S., & Cottrell, G. W. (2017). Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7272-7281.

[17]. Rueda, F. M., & Fink, G. A. (2018). Learning Attribute Representation for Human Activity Recognition. 2018 24th International Conference on Pattern Recognition (ICPR).

[18]. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable Person Re-identification: A Benchmark. IEEE International Conference on Computer Vision (ICCV).

[19]. Li, W., Zhao, R., & Wang, X. (2013). Human Reidentification with Transferred Metric Learning. Computer Vision – ACCV 2012 Lecture Notes in Computer Science, 31-44.

[20]. Millar, M. (2019). Review of Current Methods for Re-Identification in Computer Vision. Open Science Journal, 4(1).

[21]. Antipov, G., Berrani, S., & Dugelay, J. (2016). Minimalistic CNN-based ensemble model for gender prediction from face images. Pattern Recognition Letters, 70, 59-65.

[22]. Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 152-159.

[23]. Zhang, M., Wang, P., & Zhang, X. (2019). Vehicle Color Recognition Using Deep Convolutional Neural Networks. Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science - AICS 2019, 236-238.

[24]. Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv:1804.02767v1.

[25]. Lemley, J., Bazrafkan, S., & Corcoran, P. (2017). Smart Augmentation Learning an Optimal Data Augmentation Strategy. IEEE Access, 5, 5858–5869.

AUTHOR PROFILE

<Author Photo>
Matthew Charles Millar
Matthew is a researcher in deep learning and computer vision. His works investigate application on applied machine learning and computer vision algorithms to improve business applications and research.