Review of Car Make & Model Recognition Systems

Zauraiz Alamgeer[1]; Sathish Kumar Selvaperumal[2]; Sophea Prum[3]; Thang Ka Fei[4]
[1,2,4] School of Engineering, Asia Pacific University of Technology & Innovation (APU)
Technology Park Malaysia, Bukit Jalil, Kuala Lumpur 57000 Malaysia
[2]MIMOS Bhd, Technology Park Malaysia, 57000 Kuala Lumpur, Malaysia
[1]zauraiz.alamgeer@mimos.my; [2]dr.sathish@apu.edu.my; [3]sophea.prum@mimos.my; [4]dr.thang@apu.edu.my

Abstract— **This paper focuses attention towards the review of various applications and approaches in the field on image processing up to and including recent advancement of deep learning using convolutional neural networks that can be used as tools for tackling the obstacles of Car Make and Model Recognition (CMMR) in real-world environment images. Such algorithms for CMMR system are typically designed to detect specific features in images that used to be formed by feature engineering processes and are now being replaced with deep learning. The review consists of three types of algorithms. The first set explores the traditional methods that use feature extraction to localise cars in various applications and attempt to provide solution for recognizing car characteristics with feature matching over whole images in database. The next set under consideration was deep learning since it demonstrated promising results due to automatic feature engineering although still being an area under consistent research and improvement over the past few years. This paper refers to how the deep learning systems have contributed towards successful CMMR and not a comparison of deep learning architectures. The last section of this review is focused in fine-grained classification with deep learning. This is conducted especially considering the cars that are generally built up of many different parts and identifying them based on fine-grained parts from very recent researches and whether it is a viable method for attaining better overall classification accuracy score.**

Index Terms— **Machine learning, computer vision, deep learning, convolutional neural network.**

1. Introduction

Machine learning is a subset of computer vision which is one of the computer world's fastest rising new developments. There are several applications that might be utilized for this concept such as self-driving cars, learning robots or a medical system that diagnose medical images. One application of machine learning and computer vision is CMMR. Modern cars represent a new age for mobility and means of primary transport for us on day to day basis. With other surveillance aspects of urban life taking leaps forward such as face recognition systems at airports and other security related places, similarly car identification is also considered as a step forward in advanced surveillance systems. From perspective of computer vision, car identification is considered as hierarchical identification based on make, model and the production year range for specific models in assembly.

With rapid consumer demand of new and unique cars models, for each production year has leads to cars manufacturing having very large quantities of varying shapes and sizes. This altogether yields appearance differences in unlimited poses that demand today's car recognition algorithms to be very robust in terms of outside conditions, deformities and occlusions. In modern design language, cars tend to have distinctive properties such light styles, seats, configuration options and whether its sports model or economy model, all of which are recognizable from appearance. In comparison to a human face recognition, the car classification and recognition points at inferring that if the two cars belong to same make and model rather than each person having a unique face to identify, thus making this an interesting as well as less researched challenge with the primary focus on identifying model from a single image.

Till now, many methods have been proposed to read vehicle features for CMMR. The question of multiple theoretical stances on several methods has been discussed in this article and examined that how the recent and prominent researches have progressed. The three major approaches were either to perform CMMR with state of the art image classification algorithms by considering human feature engineering with localisation, or direct image recognition process with automatic feature engineering on the whole image without localization based on deep learning for detecting a known car in image individually based on its classification and lastly the fine-grained classification approach that combined feature engineering, localization automatically of sub-parts and yet their challenges faced so far in the state of the art.

2. Literature Review

2.1. Image Classification Algorithms for CMMR

Starting from the most famous classification algorithms that are still used today as building blocks of modern feature engineering tools, CMMR had been a challenge for many to achieve and the review begins with

the most basic CMMR approaches to realise the progress that have been made today.

Figure 1: Feature Matching (Cheung and Chu, 2008)

Initially Cheung & Chu (2008) proposed the methodology for CMMR is by defining interest points on cars for matching features in two images using Scale-Invariant Feature Transformation (SIFT) algorithm. Figure 1 shows two images being feature matched that represented by red lines where the points of interest were plotted on two images between test and training image and then based on geometry, the points that were in same location on the image, represented interest points on cars. The points were called inliers and the models in dataset matching to the most number of inliers that were same in test image inliers represented the classified car match. RANSAC model was used to determine symmetrical points in images that ensured that the points of interest belong to car. The disadvantage however, was that the recognition worked at same angle as dataset only.

According to Chen et al., (2015), although SIFT approach is a common feature extraction algorithm but it can be slow for real-time and some vehicles also have similar shapes even if manufactured by different manufacturers thus leading to inaccuracies. Therefore, an enhancement to Speeded Up Robust Features (SURF) algorithm initially proposed by Bay, Tuytelaars & Gool (2008) was introduced as Symmetrical-SURF. This helped to form region of interest on the axis of highest symmetry in image and detect vehicles comfortably in noisy environments such as roads. A grid was formed within the bounding box, this extracted the features based on boxes inside grid and Support Vector Machine (SVM) classifier was used for feature classification. Each of the grid within the bounding box was used independently to represent certain features of the car which increased accuracy. The grid helped in a manner that of one part of car was hidden behind a person or another car, even then the remaining grids contributed to feature extraction and helped classifier in making identification of car. This yielded 98.48% overall classification accuracy with 2864 training images and was tolerable for +-20 degree angles. Similarly, Emami, Fathi & Raahemifar (2014) also proposed CMMR from the back of the vehicle like Cheung & Chu (2008), but recently the trend had evolved to determine the car location in image symmetry from number plate rather than the entire car itself which was not accurate always due to mechanical deformities in some cases. The scale of car was predicted from number plate once detected. Using Hue Saturation Value (HSV) colour detection, the red colour was detected to represent taillight. The Region of Interest (ROI) was defined by taillights, badge and bumper. Location of number plate with reference to taillights determined the class of vehicle such as truck, car or van. The classification reduced choices of dataset resulting in faster classification. Features of taillights such as orientation, height/width and equidistance were used with Sobel edges algorithm. The k-Nearest Neighbour (k-NN) classifier was used and performed 96.3% overall classification accuracy on 280 test images on multiple camera angles but suffered with 53.1% accuracy in night time.

Moreover, Yang (2013) proposed an adaptive Harris corner detection and recognition method to identify car makes and models based on the front of the car. In this method, the number plate of car was considered as reference point to detect symmetrical features in image using Harris corner detection algorithm. This algorithm detected differences in pixels that translated to corners or edges of objects in given image. With an assumption that car logo is placed directly above the car number plate so the position of logo was estimated from refence of number plate and the Adaptive-Harris corner algorithm was developed to extract logo features which were to be later classified with SVM. With adaptive method, the threshold of Harris corner detection was increased or decreased based on light conditions in image, since too dark images would create low gradient pixels and cause unwanted corner detections. System utilised Graphics Processing Unit (GPU) for faster speeds and was tested on 1096 images consisting of 12 models moving at speeds of 20km/h and the overall classification accuracy determined was 99.5%. The limitation for this system was that it can only determine make of the vehicle from logo. Along with the different approaches commenced by various researchers on CMMR, Deep Learning had immensely improved accuracies due to GPU enabled training with NVidia CUDA capabilities since 2012. Until then the overall classification accuracies were ranging 70% on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) and with deep learning the percentage increased. The following researches used deep learning architectures.

Henceforth, when AlexNet was proposed by Krizhevshy, Sutskever & Hinton (2012), it broke all previous ILSVRC records which is basically an international image classification competition held

annually. Till now all researches mentioned required manual feature engineering and definition which was difficult on large datasets such as ImageNet dataset which consists of 1.2 million images and 1000 classes which has today become a default dataset to benchmark and compare classification accuracies by many authors. Deep Learning model possess capabilities of learning features automatically instead of engineering manually. The architecture consisted of 5 convolution layers and 3 fully connected layers. Novelty of this model consisted on using Rectified Linearization Unit (ReLU) for the activation of a neuron. Today all modern architectures use ReLU activation. This allows max function activation of neuron rather than non-linear activation which was done exponentially earlier. This property allowed neurons to fire activation even from small example trainings. AlexNet used SoftMax classifier in its fully connected layers. Each convolution layer consisted of kernel filters that automatically activated when their required features were detected in image and contributed to weights of the network training. Upon testing phase, the weights were used to find the learnt features in new unseen image and then perform the classification with fully connected layers that compared the detected weights to its trained dataset weights then give prediction. Model performed 15.3% top-5 error in ILSVRC.

With this advancement, Yang et al., (2015) realised the potential of deep learning and its application to CMMR. It was stated that deep learning needs huge dataset of images and the lack of huge training sets for deep learning hindered success. So, Yang et al., (2015) proposed a Large-Scale Car Dataset which was made to be sufficient for fine-grained classification also. The dataset when trained of deep learning architectures, was found to provide CMMR capabilities to AlexNet and similar networks with interesting findings. It was found through conducted tests that models trained with specific car parts and especially tail lights gave higher accuracies rather than training cars overall for classifications and during fine grained it was found that front and back-side poses were the most reliable angles for training networks. Verifications were made by visualizing neuron activation in last fully connected layers of networks. Total of 44,481 images having 281 car models were used in testing models like AlexNet GoogLeNet and Overfeat for surveillance camera view. Front view was found to be best suited for such camera since 98% accuracy in various weather conditions was acheived.

Szegedy et al., (2015) proposed a new model with a very deep architecture as target. The model was constructed with a network-in-network approach. It was designed to use maximum system resources and consisted of 22 layers as compared to 8 layers proposed by Krizhevsky, Sutskever & Hinton (2012) but used ReLU for activations. The norm was to increase number of layers in deep architectures to increase accuracy by in GoogLeNet, 1x1 convolution layers were added in each network and reduce the dimensionality of the network while increasing the width of network. The width of network is the number of units at a level in architecture. GoogLeNet saved computation by looking for low-level cues first like colour or texture of object and then form bounding box internally on the expected region with high probability of containing trained object and then only performing full convolution network to form ensemble of predictions. This saved from unnecessary computations. Model performed 6.7% top-5 error in ILSVRC and showed that it could be trained to perform CMMR.

2.2. **Multiple Object Detection in Image Classification**

At this point it is was affirmative that Deep Learning is capable of CMMR but since deep learning classifies an image as a whole, detecting parts of image specifically in scenarios such as detecting a car on city streets while neural network training for the same car was trained with background as country side roads. Although it is the same car and it should recognize, but notice the backgrounds are different. This caused the deep learning algorithms to perform poorly. This revelation required deep learning to recognize and learn only the car specifically and not the background in image. For this, segmentation algorithms were studied.

According to Erhan et al., (2013), Scalable Object Detection using deep neural networks was proposed after that it was observed that although deep learning had achieved very good results on benchmark datasets such as ImageNet but it still lacked the power to detect multiple objects even if same two objects in a single frame of image. This research proposed a method to individually identify objects using neural network as well but form bounding boxes on recognizable objects so that individual confidence scores can be obtained. "DeepMultiBox" algorithm was proposed that calculated confidence scores individually of each category in image for classification. Although it was not as accurate as GoogLeNet but achieved results in metric of classification as well as detection on ILSVRC-2012.

Similarly, Girshick et al., (2014) proposed Region-based Convolutional Neural Network (R-CNN). Figure 2 represented precisely how the concept of object detection was implemented in computer vision where objects were localised individually in an image and classified independently. The methodology was set to generate about 2000 proposals of regions in the input image using CNN and extract fixed length feature vectors of each

category. This approach is efficient in computation and allows each category to have independent aspect ratio. After the regions were proposed and fixed length vectors were extracted, then each region was classified with SVM specific to its own category. Object detection was designed in three modules. Firstly, the category specific regions proposals were generated and defined the candidates to be used for detection. The method used to propose regions of interest was "selective search" algorithm that proposes regions in image for segmentation based on appearance consistency. The next module was CNN, which extracted feature vectors of fixed lengths with respect to each proposed region. The proposed region was wrapped in bounding box. The final module classified the vector with specific to class linear SVMs. An overlap threshold was used in case that the object detected had been overlapped with something else it would still be detected in the bounded region.

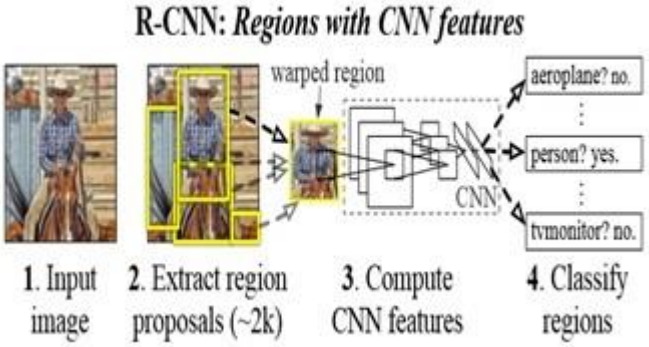

Figure 2: R-CNN (Girshick et al., 2014)

Furthermore, Ren et al., (2016) proposed Faster Region-based Convolutional Neural Network (FR-CNN); an enhanced model that took contributions such as R-CNN to further improve the detection times while improving accuracy. Although many models rely on inexpensive methods to identify the object before classification but even then, some models like Selective search like used by Girshick et al., (2014), take approximately 2 seconds on an image to detect the objects. It was faster because it utilised the same layers for both detection and classification as well. Proposals of region were generated in sliding window method and defined spatial positions of objects. Since the positions were relative so it made the system scale invariant as well. To calculate loss function, random spatial positions known as anchors were used and make predictions on detections. System scored 73.2% on COCO dataset for image classification and 42.7% mean-Average Precision (mAP) while it performed 76.1% (mAP) on PASCAL VOC 2007 (training sets from COCO +VOC 2007 + VOC 2012) as a measure for detection score.

Carrying on, Redmon et al., (2016) proposed a new architecture called You Only Look Once (YOLO). The purpose behind this system was to propose a more unified architecture. This used a single convolutional neural network and represented the bounding boxes as well but by using the entire image altogether as a whole. This was achieved by dividing the image into blocks and access them using regression technique. The architecture could use any of the CNN models like AlexNet or GoogLeNet but the approach of application was novel in YOLO that it enhanced the detection in terms of allowing such networks to perform classification and detection as well. This was made possible by novel method of regress the image down to a 7x7 tensor. The output of the network resulted in a tensor that had spatial bins defining the spatial locations of each bin. Bins contained information of class and bounding box corner coordinates. This enabled many objects to be detected by the network at the same time and in a much faster time. A drawback was that since the network formed widow being 7x7 so it was very small for certain applications and lacked in detecting tiny objects or those that were too close each other such as birds in flock due to spatial bins approach. YOLO performed in real-time speed and scored 63.4% Map in PASCAL dataset.

Also Zhou et al., (2016) addressed vehicle recognition problems with a combination of multiple approaches each targeted towards a certain short coming of deep learning. Since deep learning itself is not capable of detection within image so YOLO as proposed by [9] was used for performing detection task where it identified cars. The contribution here was that YOLO was modified to increase its grid size from 7x7 to 11x11 with a probability of only 1 class that was detecting cars in general. It was observed that of classes were increased to 2 so YOLO accuracy decreased. This showed that YOLO lacked classification capabilities. Although the system R-CNN proposed by Girshick et al., (2014) had better classification accuracy but YOLO has real-time speed thus making it more practical. For the classification part, architecture proposed by Krizhevsky, Sutskever & Hinton (2012) was used to extract low level features like Gabor filters and colours. The higher layers performed classification related extraction. The network was modified from second last layer onwards to get feature vector and use SVM classifier for classification of vehicle make and model. Testing on 987 images attained overall accuracy of 93.3% in identification and 83.3% in classification.

In a different method Wang et al., (2010) proposed object detection in image using Spatial Pyramid Matching (SPM) that detected spatial distance in between the objects detected in image. Features were extracted with

SIFT, then Locality-constrained Linear Coding (LLC) was used as a kernel to determine the scarcity of features and determine their spatial locations in image which was used by SVM classifier to improve accuracy of traditional SIFT. This resulted in mAP score of 59.3% in PASCAL VOC 2007. Followed by, Krause et al., (2016) then improved upon SPM-LLC approach by representing the data in 3D for cars unlike Wang et al., (2010) that only used 2D representation in general cases. The 3D helped to capture not only features but also shape of the car which outperformed 2D representation. Method used was to combine LLC with spatial pooling of Bubble Bank (BB) alongside SPM pooling and extract fine grained details of different car parts with help of deep learning. BB had advantage that it combined hundreds of extracted features together to represent sets of car parts which were then classified.

2.3. **Fine-Grained Classification**

Convolutional neural networks so far had proved to be successful in categorization but where there the differences amongst sub-classes are concerned, there has been very little research done on that.

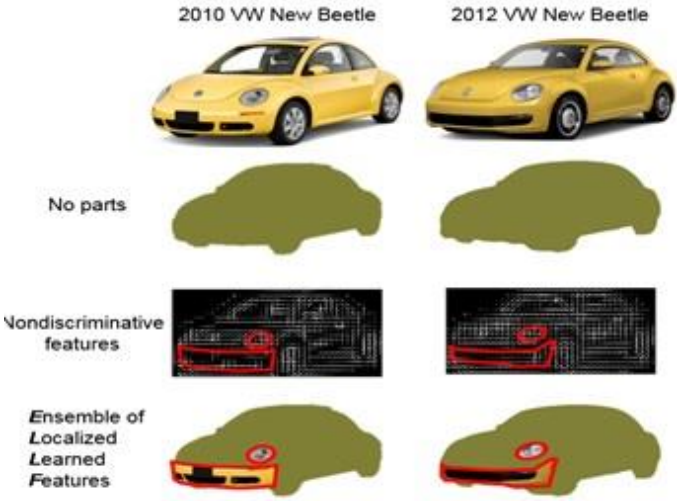

Figure 3: Fine-Grained Classification (Krause et al., 2014)

As seen in Figure 3 that which further justifies fine grained classification that although the two cars are different but if considered using overall shape or figure of the car then there are certain details that are overlooked since the property of deep learning convolutional neural networks is to generalise features. In case of car model differentiation at a level of same make but different years then the details are very much similar and have minor difference for which fine grained classification was most helpful and it was observed as Krause et al., (2014) targeted deep learning work towards car make and model recognition between similar models using fine-grained detail and proposed Ensemble of Localized Learned Features (ELLF). The system was divided into two parts where first feature learning was designed and then part discovery. AlexNet was modified to account for much smaller images than it was originally designed for. The model was redesigned to 2 convolution layers and 3 fully connected layers. The smaller size was made to avoid overfitting. The novelty in this research was that it performed unsupervised learning for the network and that meant that no human annotation was needed manually the problem with good networks is that it need large amounts of training data and data require humans to label the data for providing ground truths. With this research, the key parts of the car in image were automatically recognized from low-level cues. The parts were chosen based on highest Histogram of Oriented Gradients (HOG) energies and taken randomly as sample. Then the similar looking parts were scanned in the entire training database. The part was defined by finding the common geometric position of each part. This required the entire dataset and testing image to have similar pose. For example, if side mirrors are on side of the car from front view then all those detected parts in that particular location would be treated and labelled as side mirrors. Repeating the procedure formed an ensemble of localised features that were considered to be fine grained parts of car for training. SVM was used classify the detected parts. It was observed that randomizing the parts detection yielded better results and contributed to a more robust system. The average classification accuracy was 73.9% on car-197 (Wang et al., 2016) dataset.

Bay, Tuytelaars & Gool (2008) then stated that sub-categories in an image can also be labelled in unsupervised manner with a segmentation and alignment directly. With this method, the parts were generated with segmentation which caused the subject in image to be focused and removed background. Alignment over all the dataset was still necessary as with earlier research (Krause et al., 2014). Alignment helped to generalise the localization of parts and made patch extraction more refined. Once all the parts were extracted then k-means clustering algorithm was used to cluster all the similar looking patches together. Since not all the patches are useful to give discrimination such as tires in all cars look similar or as in this research the bird's dataset was used to eyes are similar many species so the max-margin template selection algorithm was used that utilised alignment and pose. This was an important aspect since it can increase or decrease effectiveness of entire patch extraction process. This was much like a one-class SVM in which the sparse weights of classifier were chosen to train for only a positive class and no negative class. This resulted in SVM decision values to be applied

using cross-validation for each part and provide its class. The parts were all individually trained as positive classes and used with a modified joint configuration point scores for normalising log-probabilities of each patch and applied score globally with respect to other patches in the given image thus enabling to tell if a patch had high classifier score and was discriminative enough for a specific class. The accuracy of the system was tested in cars-197 dataset that resulted in 92.8% classification accuracy of cars with only annotation of bounding box of cars given during training.

Similarly, Wang et al., (2016) too stated that for cars, the classification with deep learning suffered at discriminating sub-categories in images. It was believed that subtle differences lay in the localized areas of cars that pointed the challenge to determine discriminative regions in a class. The goal was similar to earlier research that first find the discriminative regions and then find a way to find most discriminative amongst the rest. The methodology was to form patches of triplets per image and let them define the property of features on three different geometric position on the image. For that, mining system was used on entire dataset to find discriminative patches. The constraints such as geometric and order were used to make triplets on every image. All the cars were aligned similar in dataset and the discriminative score was determined with ratio with in-class variations from each other as a whole. The highest response of triplets was taken from mid-level responses of image features of the patches. For classification, the response of triplets per image was converted to a one descriptor representation per image and then used with SVM to classify. Testing on cars-197 got 92.5% accuracy of classification.

3. Results and Discussion

With careful review of existing systems, tables were drawn to summarize their performances and how the systems compete to each other. The tables below summarize the results and compare system wise with first comparison between standard image classification followed by localization algorithms and lastly fine grained classification systems.

Table 1: Comparison of Methodologies for image classifications

Year	Researcher	Methodology	Results
2015	Yang et al.,	Large dataset for fine-grained classification	98% (281 models tested)
2015	Szegedy et al.,	GoogLeNet (Inception Module)	6.67% top-5 error rate ILSVRC (2014)
2014	Emami, Fathi & Raahemifar	Tail light detection with k-NN	96.3%
2013	Yang et al.,	Adaptive HARRIS Corner Detection	99% (12 Models)
2013	Hsieh, Chen & Cheng	SURF with SVM	98.4% (Small Dataset)
2012	Krizhevsky et al.,	AlexNet	15.3% top-5 error rate ILSVRC (2012)

Results in Table 1 showed that other than deep learning the overall classification accuracies although attained high scores such as 99% in [14] but it is also to be considered that it consisted of testing on only 12 models. Algorithms consisting of deep learning proved capable of performing well on ILSVRC dataset with top-5 error rate as standard metric of evaluation. It was understood that overall comparison is not fair due to differences in datasets used for testing but standard benchmark was preferred and considered more challenging.

This leads to confirmation that deep learning outperforms all other state of the art for image classification but remained short on localization for which it was combined with object detection algorithms thereon to provide more real-life oriented applications where objects such as cars are found in noisy backgrounds. Then the question remained that which algorithm performed best localization. Results comparison is shown in Table 2.

Table 2: Comparison of Methodologies for image object detections

Year	Researcher	Methodology	Test Dataset	Detection Results
2016	Redmon et al.,	YOLO with VGG-16 architecture	PASCAL VOC 2007	66.4% mAP with Training VOC '07 + VOC '12
2016	Zhou et al.,	Modified YOLO + Modified AlexNet	Detection 987 Images	93.3% Detection Precision
2016	Ren et al.,	Faster R-CNN with VGG-16 architecture	PASCAL VOC 2007	70.4% mAP with Training VOC '07 + VOC '12
2013	Erhan et al.,	DeepMultiBox	PASCAL VOC 2007	29% mAP
2014	Girshick et al.,	R-CNN	ILSVRC 2013	29.7% mAP

2010	Wang et al.,	LLC SIFT with SVM	PASCAL VOC 2007	59.3% AP

Results viewed that over the years a trend towards improvement was seen and standard metric of evaluation was mAP. For fair comparisons, the configurations of all the methods were kept as similar as possible amongst which [9] and [10] proved to be most efficient and performed highest on standard benchmark dataset PASCAL VOC 2007. Although FR-CNN had higher mAP but in real life YOLO had better frame rate 7 FPS vs 21 FPS respectfully which makes it more suitable for real-time applications.

Table 3: Comparison of Methodologies specific for Fine-Grain Approaches

Year	Researcher	Methodology	Test Dataset	Results (Average Classification Accuracy
2016	Krause et al.,	k-means and 1-class SVM for refining discrimination	Car-197	92.8%
2016	Wang et al.,	Extract triplets of patches based on geometric constraints after aligning images	Car-197	92.5%
2014	Krause et al.,	ELLF	Car-197	73.9%

Only localizing and then applying classification was still prone to confusion between similar looking objects so the fine-grained classification helped in intra-class classification. In Table 3, the results are measured as overall classification accuracy score and shows comparison of latest algorithms for CMMR in fine-grained state of the art classification.

4. Conclusions

From the literature review conducted above it was observed that with each new research, more advancement has been made towards image recognition tasks. It was observed that initially the trend for CMMR was to localize the car in the image by using number plates or car taillights as reference points and estimating the car positions. Then image feature extraction was engineered by hand defined filters which was not always perfect and did not perform robust to multiple cars in an image. Then Deep Learning was rediscovered due to GPU capabilities enhancing the training times that once took weeks to train now took days only. With fast deep learning capabilities, many new algorithms were discovered such as AlexNet and GoogLeNet. Although deep learning provided good image classification scores but from reviews it was seen that when used standalone, it suffered in multiple instances such as two or three cars in an image but such architectures became building blocks of modern systems and are still used today even in the latest of researches so they hold vital and critical contribution to the subject. Using those architectures, new researches pointed towards image segmentation algorithms that enabled cars to be detected with bounding boxes in an image with the high accuracy of deep learning.

Algorithms such as YOLO or R-CNN, performed image detection and then classification of the detected object altogether. This performed high classification scores when then in the cutting-edge technology, the CMMR had been taken to next extremes by using fine-grained classifications that performed by pointing discriminative parts of the image in an unsupervised learning fashion. The latest work with fine-grained classification, pointed critical parts of cars body that helped to identify the car by its special characteristics. For instance, comparing the highest scoring localization method, R-CNN with fine grained method introduced by Krause et al., (2016), that even if detection is slightly inaccurate but fine grain still classifies correctly with localization because R-CNN had to treat the entire object to be detected in image as one bounding box hindering it differentiate in fine detail of objects like car bumper lights and shapes.

The only limitation in today's technology is that it requires pose normalisation that all the images of cars should have similar pose and must be aligned because the patch discrimination is performed based on geometric constraints on image. In future researches, the limitation of alignment and pose has to be considered and methods of finding the most discriminative patches amongst all patches detected in different car models have to be researched without limiting the pose. For example, if different poses are used and there are no geometric constraints then the wheels of car in an individual image might be considered as a distinctive feature of the car but as a whole when applied to the entire dataset in unsupervised learning, all the images would have wheels and detected in all the cars and the patch would no longer be discriminative as a whole since all the cars would have it and that's why geometric constraints and similar poses are used in all the state of the arts. In future if such challenges are tackled then CMMR can be performed

with very high accuracy from almost any angle without human manual annotation of fine grained training.

References

Bay, H., Tuytelaars, T. and Gool, L. (2008) SURF: Speeded Up Robust Features. Computer Vision and Image Understanding. 110(3). p. 346-359.

Chen, L., Hsieh, J., Yan, Y. and Chen, D. (2015) Vehicle make and model recognition using sparse representation and symmetrical SURFs. Pattern Recognition. 48(6). p. 1979-1998.

Cheung, S. and Chu, A. (2008) Make and Model Recognition of Cars. Projects in Vision and Learning.

Emami, H., Fathi, M. and Raahemifar, K. (2014) Real Time Vehicle Make and Model Recognition Based on Hierarchical Classification. International Journal of Machine Learning and Computing. 4(2). p. 142-145.

Erhan, D., Szegedy, C., Toshev, A. and Anguelov, D. (2013) Scalable Object Detection Using Neural Networks.

Girshick, R., Donahue,J., Darrell, T. and Malik, J. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation.

Krause, J., Gebru, T., Deng, J., Li, L. and Fei-Fei, L. (2014) Learning Features and Parts for Fine-Grained Recognition.

Krause, J., Jin, H., Yang, J. and Fei-Fei, L. (2016) Fine-Grained Recognition without Part Annotations.

Krause, J., Stark, M., Deng, J. and Fei-Fei, L. (2013) *3d object representations for fine-grained categorization*. In 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13).

Krizhevsky, A., Sutskever, I. and Hinton, G., E. (2012) ImageNet Classification with Deep Constitutional Neural Networks.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection.

Ren, S., He, K., Girshick, R. and Sun, J. (2016) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.

Szegedy, C., Liu, W., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) Going Deeper with Convolutions.

Wang, J., Yang, J., Yu, K., Huang, T. and Gong,Y. (2010) Locality-constrained Linear Coding for Image Classification.

Wang, Y., Choi, J., Morariu, V., I. and Davis, L., S. (2016) Mining Discriminative Triplets of Patches for Fine-Grained Classification.

Yang, H., Zhai, L., Li, L., Liu, Z., Luo, Y., Wang, Y., Lai, H. and Guan, H. (2013) An Efficient Vehicle Model Recognition Method. Journal of Software. 8(8).

Yang, L., Luo, P., Loy, C. and Tang, X. (2015) A Large-Scale Car Dataset for Fine-Grained Categorization and Verification.

Zhou, Y., Nejati, H., Do, T., Cheung, N. and Cheah, L. (2016) Image-based Vehicle Analysis using Deep Neural Network: A Systematic Study.