SFNet: Singular Feature-based Classification Network for Low-resolution Image

Seongmin Kim, Youlkyeong Lee, Kanghyun Jo
School of Electrical Engineering
Dept. of Electrical, Electronic and Computer Engineering
University of Ulsan, Ulsan, Korea
asdfhdsa1234@mail.ulsan.ac.kr, yklee@islab.ulsan.ac.kr, acejo@ulsan.ac.kr

***Abstract*—This paper proposes a model to improve the classification performance of low-resolution images by training with high-resolution images. Using a pre-trained model from a published open image dataset can quickly generate a high-performance vision system. Open image datasets generally have clear shapes of objects. However, trained models show bad classification accuracy on low-resolution objects in wide images such as drone-shot imagery. Low-resolution objects have relatively ambiguous low-level features to high-resolution objects. This ambiguity can limit the classification ability of trained models. To address this issue, this work proposed SFNet with Singular Feature Extract Module (SFEM) to improve predict performance. SFEM extracts the principal components of feature maps by performing Singular Value Decomposition (SVD) and concatenates them behind the convolution output. The performance was the highest when using the singular value rank-1 matrices within top 4. Pre-trained original MobileNet V1 showed 17.15% accuracy on 65x65 drone-shot object. The proposed model achieved 38.47% accuracy on same test dataset.**

***Index Terms*—Drone-shot Imagery, Image Classification, Singular Value Decomposition**

I. Introduction

With the development of a new type of mobility technologies, many types of drone research are conducted. Among them, the drone surveillance systems are also receiving attention [1] [2]. Surveillance systems should use wide-view images to reduce blind spots. In such cases, a drone is a suitable option for surveillance. Drones can fly at high altitudes to capture the wide-view images. Drone surveillance systems are required several technologies, including automatic control, data communication, and computer vision. Especially, object detection is a very important method in the computer vision of the drone surveillance system. Object detection is a task that simultaneously performs localization to find out where an object is located in an image and classification to find out what kind of object. Object detection usually uses pre-trained classification model. Most of pre-trained model is trained by high-resolution object image. But drone-shot imagery has very low-resolution object due to the high flight altitude of the drone, such as Fig. 1. This resolution gap causes decreasing the accuracy of classification. Because the low level features(e.g. edge, corner) of low-resolution are relatively ambiguous than high-resolution image. The differences of low level feature are shown on Fig. 2. This phenomenon makes it difficult to judge the class of the feature map. Therefore, this paper proposes a network with module that extracts unique features of objects through Singular Value Decomposition (SVD) and corrects ambiguous low-level feature map. Using the proposed module, classifiers will be less affected by the resolution of objects.

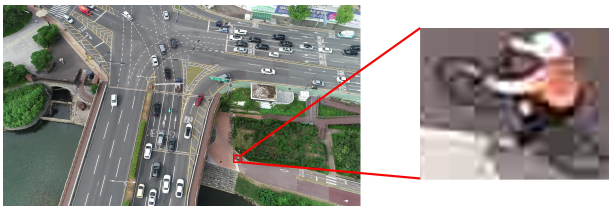
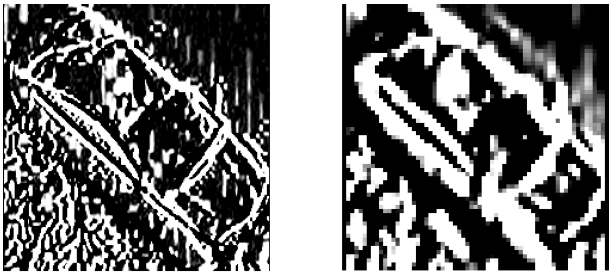


Fig. 1: A drone-shot imagery and its cropped object.



(a) **(b)**

Fig. 2: Feature maps are obtained by applying the Sobel x filter to an object at different resolutions. (a) a feature map for a 224×224 (b) a feature map for a 65×65.

II. Related work

A. Singular Value Decomposition

Singular Value Decomposition (SVD) is one of the most significant theories in linear algebra. SVD solves two limitations of eigendecomposition.

1) Only square matrix can use eigendecompostion.
2) Only symmetric matrix can be expressed the sum of rank-1 matrix by eigendecomposition.

Before explaining SVD, this paper first explains eigendecomposition. Eigendecomposition is one of the matrix decomposition methods. Using eigendecomposition on a matrix can diagonalize the matrix. The strongest ability of eigendecomposition is analyzing unique components of matrix. A matrix

can be decomposed into rank-1 matrices. Each rank-1 matrix is composed of orthonormal eigenvector and one eigenvalue. However, eigendecomposition is suitable for a square matrix. Furthermore, analyzing unique components can be only applied on symmetric matrices. SVD is usually used in general mathematical methods case, especially for image processing tasks. Because a number of images are not in symmetric form. SVD is similar method with eigendecomposition. It can be applied on rectangular matrix. The matrix can be decomposed Left-orthogonal matrix $\mathbf{U}$, Singular value diagonal matrix $\mathbf{\Sigma}$ and Right-orthogonal matrix $\mathbf{V}$ by SVD. One operation is not enough to get two different orthogonal matrices. A little trick is used. It is applying eigendecomposition on $\mathbf{A^T A}$ and $\mathbf{A A^T}$. Because $\mathbf{A^T A}$ and $\mathbf{A A^T}$ are symmetric matrices even though $\mathbf{A}$ is a rectangular matrix. $\mathbf{U}$ is eigenvector matrix of $\mathbf{A A^T}$ and $\mathbf{V}$ is eigenvector matrix of $\mathbf{A^T A}$. $\mathbf{\Sigma}$ is square root of eigenvalue diagonal matrix of $\mathbf{A^T A}$ and $\mathbf{A A^T}$. SVD can be applied on many image processing task like image compressing [3].

B. Low-resolution Classification on Pre-trained Model

Image classification models developed rapidly with the advent of CNN. Classification models with high performance, such as VGGNet [4], GoogLeNet [5], and ResNet [6], were announced. These models are used as the backbone of various deep learning network for computer vision. High accuracy classifier takes a long time to learn. Because the number of parameters is too large. Many backbone classifiers are pre-trained. The pre-trained models are usually learned by high-resolution images. However, it shows very low classification performance on low-resolution images, such as objects in drone-shot imagery. The predict results are shown in Table 1. It can be observed that the model pre-trained on high-resolution images has lower accuracy on low-resolution test images. To solve this issue, this paper proposes a novel feature extract module for low-resolution.

TABLE 1: Prediction results on drone-shot imagery datasets with different resolutions using MobileNet V1 [7]. It was pre-trained by 224×224 resolution images.

Test Dataset Resolution / Classes of Test Dataset	224×224	65×65
Car	93.4%	1.90%
House	99.9%	53.7%
Person	99.5%	1.3%
Pole	99.8%	22.50%
Tree	99.9%	0%
Truck	92.6%	23.5%
Average	97.52%	17.15%

III. Proposed Method

A. SFNet Architecture

Fig. 4 expresses structure of SFNet. Model is based on MobileNet V1 [7]. There are 14 convolution layers in model. Network uses two types of convolution sequence. They are shown in Fig. 3. First is general convolution. Second one is depthwise separable convolution [7]. This model uses ReLU [8] and Batch Normalization [9] after convolution. First general convolution layer and second depthwise separable convolution layer [7] extract low level feature about input image and feature map. When the low-resolution image and feature map include in general convolution layer and depthwise separable convolution layer [7], ambiguous features are extracted. To accurately recognize features of object, unique information is necessary for the object. This paper adds Singular Feature Extract Module (SFEM) between general convolution layer and depthwise separable convolution layer [7] input and output terminal. SFEM can extract principal component of an image with SVD. Outputs of SFEM are concatenated with convolution outputs and entered next convolution layer.

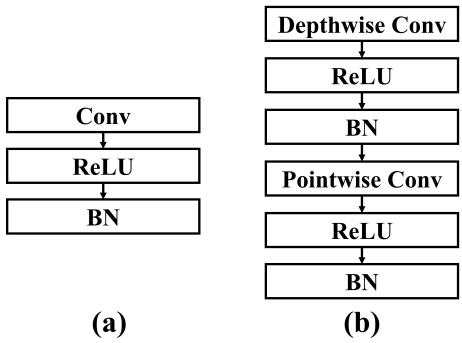


Fig. 3: Block diagram of convolution operation sequence in this model. (a) sequence of general convolution (b) sequence of depthwise separable convolution

B. Image Decomposition with SVD

If A is $\mathbb{R}^{M\times N}$ full-column rank rectangular matrix, SVD is represented as Eq. 1. This representation is very similar to eigendecomposition but SVD can be applied on rectangular matrix.

$$\mathbf{A} = \mathbf{U\Sigma V^T} = \sigma_1 \mathbf{u}_1 {\mathbf{v}_1}^{\mathbf{T}} + \sigma_2 \mathbf{u}_2 {\mathbf{v}_2}^{\mathbf{T}} + \cdots + \sigma_n \mathbf{u}_n {\mathbf{v}_n}^{\mathbf{T}} \quad (1)$$

Where $\mathbf{U}$, $\mathbf{V}$ are orthogonal matrices which has left-singular vector $\mathbf{u}$, right-singular vector $\mathbf{v}$. $\mathbf{\Sigma}$ is diagonal matrix which has singular value σ. Let a 1-channel image as $H \times W$ matrix $\mathbf{A}$. If applying SVD to grayscale image $\mathbf{A}$, the image can be expressed as the sum of rank-1 matrices, as described in Fig. 5 [3]. Let flattening a grayscale image $\mathbf{A}$ of $\mathbb{R}^{M\times M}$ denotes vector $\mathbf{a} \in \mathbb{R}^{M^2}$. Let vector $\mathbf{q}_i$ denote flattening a i-th rank-1 matrix. $\mathbf{a}$ can be expressed as linear combination of vectors from $\mathbf{q}_1$ to $\mathbf{q}_M$. Where $\mathbf{a}, \mathbf{q}_i \in \mathbb{R}^{M^2}, i \in [1, M]$.

$$\mathbf{a} = \mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3 + \cdots + \mathbf{q}_M \quad (2)$$

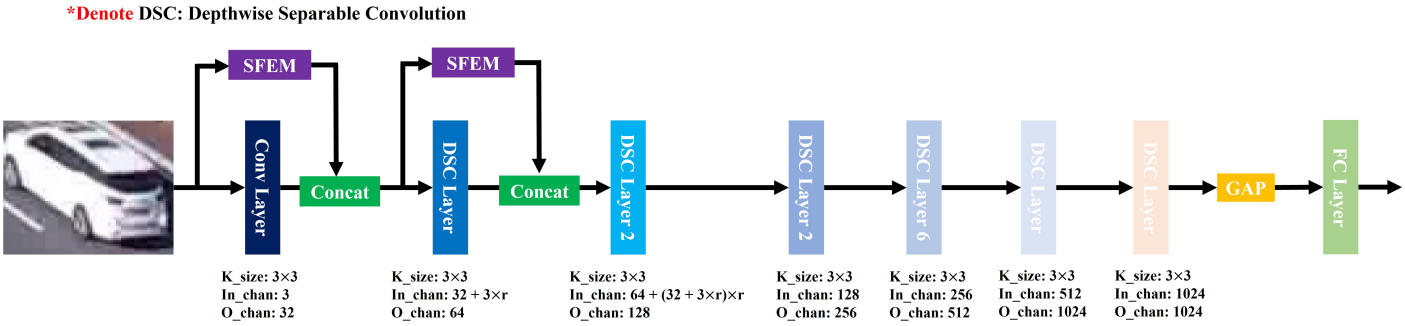


Fig. 4: Structure of SFNet. The model structure is based on MobileNet V1 [7]. As the number of channels in the convolution layer increased, lighter colors were used to express it.

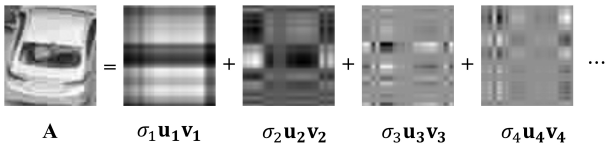


Fig. 5: Singular Value Decomposition about grayscale car image.

Eq. 2 shows that the vector $\mathbf{q}_1$ to $\mathbf{q}_M$ have linearly independent relationship. Therefore, each rank-1 matrix represents a unique and independent component of matrix $\mathbf{A}$.

C. SFEM: Singular Feature Extract Module

SFEM extracts singular features of each channel and concatenates features. First, the image or feature map is separated channel-wise in SFEM. Next the module take SVD on each channel of feature map. After SVD, principal components are generated. Where the number of principal components is as much as height of the feature map. Using all singular features can cause a large computational cost. This paper sets a hyperparameter $\mathbf{r}$. $\mathbf{r}$ is hyperparameter for selecting the number of singular feature. Principal components with large singular values are selected as a priority. After feature selecting, 4 singular features of each channel are concatenated behind input feature map. This paper selects $\mathbf{r}$=4. Because four singular features can contain enough information about the object. The Fig. 7 shows it. Using less than 4 features makes image which is difficult to understand. Merging 4 features generate relatively clear object. However, using more than 4 features to merge images results in quite similar clarity to each other.

IV. Experiment

A. Dataset

In experiment, this paper used drone-shot imagery on Ulsan and Daegu in South Korea. This paper contains whole number of images that is cropped by object detector, YOLOv5 [10]. The dataset consists of the following classes: 'Car', 'House', 'Person', 'Pole', 'Truck', and 'Tree'. Dataset has total 60,000

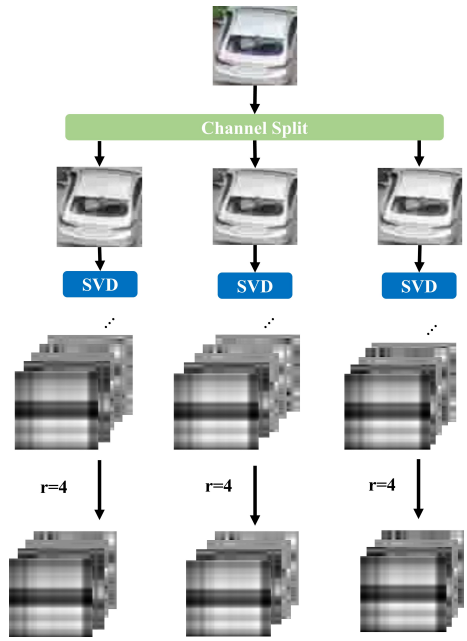


Fig. 6: Operation process of SFEM. In figure, each channel of feature map is expressed grayscale image.

images. The imageset is separated into 42,000 images for train, 12,000 images for validation, and 6,000 images for test. More detailed information on the dataset is shown in Table 2.

TABLE 2: Detail information for dataset.

Classes	Images	Train	Validation	Test
Car	10,000	7,000	2,000	1,000
House	10,000	7,000	2,000	1,000
Person	10,000	7,000	2,000	1,000
Pole	10,000	7,000	2,000	1,000
Tree	10,000	7,000	2,000	1,000
Truck	10,000	7,000	2,000	1,000
Total	**60,000**	**42,000**	**12,000**	**6,000**

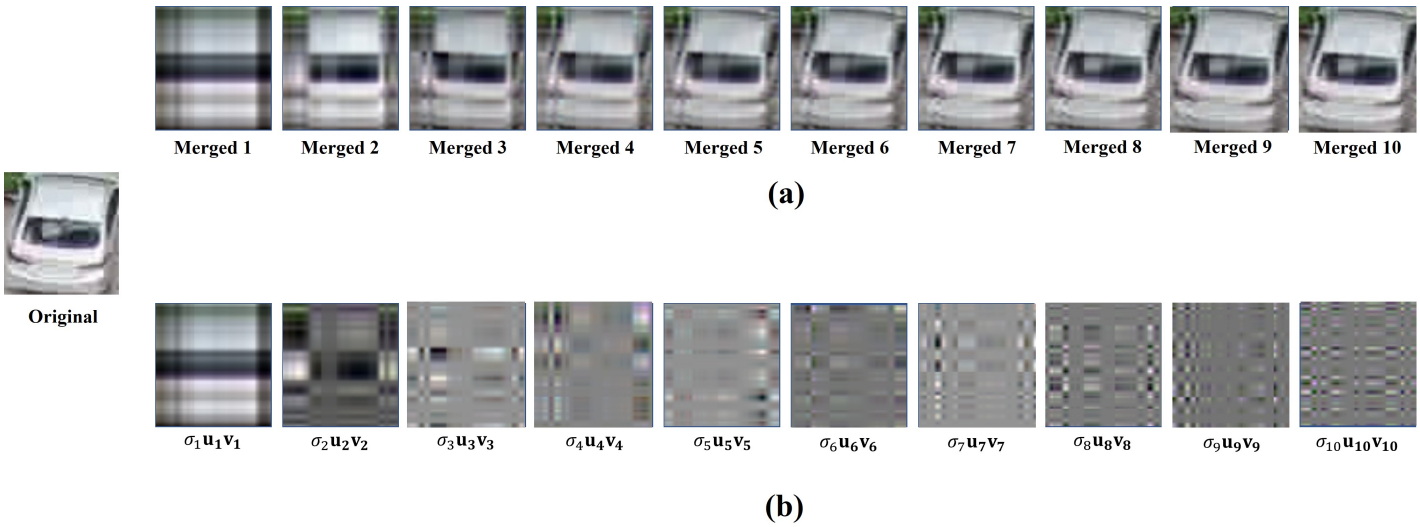


Fig. 7: Principal component analysis on car object image. Sentence (a) refers to the merged image. 'Merged n' means that 'n' principal components are used to create the merged image. (b) is principal component of car image. Each component was placed in a large singular value order.

B. Comparison

In this part, this paper is comparing the prediction accuracy of the original MobileNet V1 [7] and the proposed model on low-resolution test images. During training, the model learned with 224×224 resolution images. Before testing, the test dataset was resized to 65×65 resolution to evaluate the performance of model on low-resolution images. However, the validation dataset was not resized. Because this paper aimed to obtain pre-trained weight from high-resolution images. The detailed experimental environments are described below.

- CPU: Intel(R) Core(TM) i9-10900X
- Graphic Card: Nvidia Geforce RTX 3090 * 4EA
- Batch size: 512
- Epoch: 20
- Loss function: Cross Entropy
- Optimizer: Adam [11]

1) Comparing MobileNet V1 with proposed Network: Predict performance according to two models is described below Table 3. SFNet achieved 21.23%p increased average accuracy compare with MobileNet V1 [7]. Both MobileNet V1 [7] and SFNet show bad performance at car and person classes. Because the shape of these classes is compressed to 65x65. Therefore SFEM could not extract any singular feature.

2) Proving hypothesis about hyperparameter: To test the hypothesis regarding the value of hyperparameters, this paper trained the proposed model multiple times while changing the value of **r**. Based on the Table 4, the model achieved best classification performance when the value of **r** was set to 4.

TABLE 3: Test results on drone-shot imagery dataset with low-resolution reshaping. Each model was pre-trained by 224×224 resolution images.

Classes \ Model	MobileNet V1 [7]	SFNet(r=4)
Car	**1.90**%	0.30%
House	**53.7**%	37.80%
Person	**1.3**%	0.10%
Pole	22.50%	**53.30**%
Tree	0%	**73.80**%
Truck	23.5%	**65.50**%
Average	17.15%	**38.47**%

TABLE 4: Average accuracy of SFNet on different hyperparameter **r**.

r	1	2	3	4	5
Average Accuracy	17.50%	17.07%	30.73%	**38.47**%	29.42%

V. Conclusion

This paper proposed a novel approach for classifying low-resolution objects in drone images by leveraging high-resolution training data. The proposed model employed SVD to extract singular features for each object which were then used in an attention mechanism to improve classification accuracy. In the experimental result, the vanilla MobileNet V1 [7] achieved 17.15% average accuracy on low-resolution classification. However proposed model got 38.46% average accuracy on same resolution images. Moreover, the most significant performance improvement was achieved when attention with singular value rank-1 matrices with in top-4. In the future, proposed method is planned to explore the various classification models and investigate its effectiveness as a backbone.

ACKNOWLEDGMENT

This results was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003)

REFERENCES

[1] Youlkyeong Lee, Qing Tang, Jehwan Choi, and Kanghyun Jo. Low computational vehicle lane changing prediction using drone traffic dataset. In *2022 International Workshop on Intelligent Systems (IWIS)*, pages 1–4, 2022.

[2] Jehwan Choi and Kanghyun Jo. Attention based object classification for drone imagery. In *IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–4, 2021.

[3] H.S. Prasantha, H.L. Shashidhara, and K.N. Balasubramanya Murthy. Image compression using svd. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, volume 3, pages 143–145, 2007.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[8] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[10] Glenn R. Jocher, Alex Stoken, Jiří Borovec, NanoCode, Ayushi Chaurasia, TaoXie, Liu Changyu, Abhiram, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang, Jan Hájek, Laurentiu Diaconu, Marc, Yonghye Kwon, Oleg, wanghaoyang, Yann Defretin, Aditya Lohia, ml ah, Ben Milanko, Ben Fineran, D. P. Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham. ultralytics/yolov5: v5.0 - yolov5-p6 1280 models, aws, supervise.ly and youtube integrations. 2021.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.