sciendo

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **22**, No 1
Sofia • 2022 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2022-0007

Combination of Resnet and Spatial Pyramid Pooling for Musical Instrument Identification

Christine Dewi, Rung-Ching Chen

Department of Information Management, Chaoyang University of Technology 168, Jifeng E. Rd., Wufeng District, Taichung City 413310, Taiwan
E-mails: crching@cyut.edu.tw

***Abstract*:** *Identifying similar objects is one of the most challenging tasks in computer vision image recognition. The following musical instruments will be recognized in this study: French horn, harp, recorder, bassoon, cello, clarinet, erhu, guitar saxophone, trumpet, and violin. Numerous musical instruments are identical in size, form, and sound. Further, our works combine Resnet 50 with Spatial Pyramid Pooling* (*SPP*) *to identify musical instruments that are similar to one another. Next, the Resnet 50 and Resnet 50 SPP model evaluation performance includes the Floating-Point Operations* (*FLOPS*), *detection time, mAP, and IoU. Our work can increase the detection performance of musical instruments similar to one another. The method we propose, Resnet 50 SPP, shows the highest average accuracy of 84.64% compared to the results of previous studies.*

***Keywords*:** *Resnet 50, Resnet 50 SPP, spatial pyramid pooling, musical instruments, similar object.*

1. Introduction

An example of computer technology that is closely connected to computer vision and image processing is object detection. This technique is involved with locating instances of semantic items belonging to a certain class, such as musical instruments [1], buildings [2], people [3], traffic sign [4] or cars [5] in video and digital images [6]. Despite the widespread usage of object detectors, the effectiveness of features extracted may be inconsistent in particular situations. A good example of this is when two object classes have similar appearance to each other as seen in Fig. 1. Therefore, the detector blears the class of the object being observed, pretending that different classes of objects with comparable appearances are referred to as related object pairs [7].

The flute and clarinet have a number of features that are complimentary. Clarinet is a woodwind instrument with a single-reed mouthpiece, a cylindrical tube with a flared end, and holes stopped by keys. Flute and clarinet are both essential members of the woodwind family of instruments, and they are often used together. One of the most significant differences between clarinet and flute is the presence or

lack of reeds; flutes are reedless instruments, whilst clarinets are instruments with just a single reed. Furthermore, the clarinet is a side-blown instrument, while the flute (western concert) is an end-blown instrument. In terms of form, size, and sound, they are identical to one another. The primary difference between a cello and a violin is size. The cello is normally played from a seated position with the instrument held between the legs. The violin, in contrast, is held between the shoulder and the chin. The cello produces lower notes on the scale than the violin. Nevertheless, as shown in Figure 1, the violin, cello, and guitar all have a similar basic shape. While it is easy for humans to identify comparable musical instruments, computers have a much harder time.

(a) Bassoon (b) Recorder (c) Clarinet (d) Guitar (e) Violin (f) Cello

Fig. 1. Similar musical instruments

This article examines in-depth Convolutional Neural Network (CNN) models and features extractors, particularly Resnet 50 and Resnet 50 SPP for object recognition, as well as feature extraction method. The architecture of Resnet won the contest of ImageNet in 2015 and comprised so-called Resnet blocks [8, 9]. Instead of reading a function, the residual block barely learns the residual and is consequently preconditioned in each layer to learn mappings that are approaching to the identity function.

Our research fine-tunes them to the People Playing Musical Instrument (PPMI) dataset [10]. The PPMI dataset includes images of people interacting with various musical instruments, including twelve different instruments. Cello, bassoon, clarinet, flute, French horn, erhu, guitar, harp, recorder, trumpet, saxophone, and violin are some of the instruments on the list. It is not difficult in research papers to come across object detectors based on deep learning specifically customized to the traffic sign detection issue domain. We have struggled to locate one that evaluates many important variables, such as mAP, IoU, and detection time.

A brief overview of the contributions of this paper is as follows. First, we seek to identify objects that are very similar at the level of human vision. Second, we combine Resnet 50 with Spatial Pyramid Pooling (SPP) for the identification of musical instruments that are similar to one another. Following that, the Resnet 50 SPP model assessment covers the detection time, mAP, IoU, and Floating-Point Operations (FLOPS). In this study, we recognize between many musical instruments that are comparable to each other, including harp, bassoon, clarinet, cello, erhu, flute, recorder, French horn, saxophone, trumpet, and violin.

The remainder of this paper is organized as follows: Materials and methods are given in Section 2. Section 3 describes the experimental results and discussions. Finally, in Section 4, conclusions are made, and recommendations for further research are proposed.

2. Materials and methods

2.1. Similar object identifications

Deep learning recognition has allowed considerable advancements in the majority of object identification algorithms over the last few years [11]. Object identification is easy for humans, but it is very challenging for computers to differentiate between two objects that are almost similar in both look and function. The two-stage detector is made up of two processes that function in conjunction with one another. For starters, the detector derives suggestions for locations where things may be located in the picture by using a Region-based CNN (RCNN) technique. Later, each Region of Interest (RoI) is classified independently and then combined [12, 13].

Even though the two-stage detector has good performance, it does have some important limitations, which are as follows: Because of the two techniques requirements, long time is needed to train a model and much longer to evaluate it. For the shortest amount of forecast time, a single-stage detector is suggested. Redmon et al. [14] and Single Shot Detector (SSD) [15] are the most representative single-stage detectors. Both have only one-stage CNN architecture. Compared with two-stage detectors, single-stage detectors have fewer model parameters, they are faster, and have more competitive overall performance. Ju, Moon and Yoo [7] present an object recognition method that incorporates entropy loss to better correctly identify items with similar appearances. When entropy loss is used, the detector makes more robust predictions about the observed bounding box class, resulting in a greater probability of a good score. Additionally, it has the effect of decreasing trust erosion. As a result, the detection performance of comparable things is enhanced. In [16], global as well as local self-similarity descriptors are discussed to find the similar object detection. Similar object detection can be done using any of the two well-known techniques, i.e., global self-similarity descriptor and local self-similarity descriptor.

In our research work, we recognize many musical instruments that are comparable to each other, including the harp, bassoon, clarinet, cello, erhu, flute, recorder, French horn, saxophone, trumpet, and violin.

2.2. Spatial Pyramid Pooling (SPP)

In our research study, we recognize many musical instruments that are comparable to each other, including the harp, bassoon, clarinet, cello, erhu, flute, recorder, French horn, saxophone, trumpet, and violin. SPP [17, 18] has been shown to be much more successful than other methods in object identification tests. Consider the seriousness of the issue – this is a competition between approaches, which make use of increasingly advanced spatial modelling techniques. The picture of the spatial pyramid is divided into a series of finer grids at each level of the pyramid in order to aid in the understanding. Also, it is *commonly-known* as Spatial Pyramid Matching (SPM) [19], a development of the Bag-of-Words (BoW) model [20], which is one of the most famous and successful methods in computer vision methods. SPP has continued being an important component and superior system to win the competition in the classification [21, 22] and detection [23] before the recent ascendance of CNN.

SPP [24] provides the following advantages, which may be stated: First and foremost, SPP is capable of producing a fixed-length output regardless of the input dimension. Second, SPP utilizes multi-level spatial bins, while sliding window pooling only employs a single-window size, which is in contrast to SPP. Following that, SPP enables us to not only produce pictures from arbitrarily sized photos for testing, but also to input images with varying sizes and scales into the system during training. Aside from that, training by using variable-size photos increases the amount of variation in size and reduces overfitting. Additionally, SPP is particularly congenial to detecting objects. Deep convolutional networks are used to extract features from candidate windows in the R-CNN object identification technique, which is the most widely used object recognition approach. Furthermore, SPP may integrate characteristics produced at varied scales with the flexibility of input scales, resulting in a more powerful algorithm. CNN layers take inputs with arbitrary sizes, but they create outputs with a wide range of sizes as well.

2.3. Resnet 50 SPP

Resnet 50 which is a short form of Residual Networks is a deep learning architecture [25]. The Resnet is similar to other deep networks but it has an additional identity mapping capability. Resnet models fit a residual mapping to predict the delta needed to reach the final prediction from one layer to the next. This has shown that it can successfully address the vanishing gradient problem. Resnet 50 is characterized by a very deep network and contains 34 to 152 layers [26, 27]. This architecture can be seen in Fig. 2 being developed by researchers at Microsoft and having won the ILSVRC 2015 classification task [28]. In the Resnet model, a residual network structure is implemented.

In [29], a new TCNN (ResNet-50) with the depth of 51 convolutional layers is proposed for fault diagnosis. By combining with transfer learning, TCNN (ResNet-50) applies ResNet-50 trained on ImageNet as feature extractor for fault diagnosis.

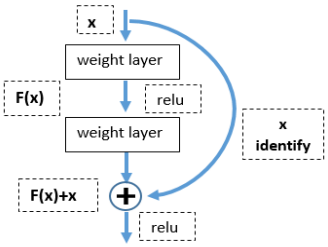


Fig. 2. Residual block

In [30] early diagnosing Alzheimer's Disease (AD) facilitates family planning and cost control. A Residual Network with 50 layers (ResNet-50) has predicted the Clinical Dementia Rating (CDR) presence and severity from Magnetic Resonance Imaging (MRI)'s (multi-class classification). Machine learning methods classify AD with high accuracy. ResNet-50 network models might help identify AD patients automatically prior to provider review. Fig. 3 describes our Resnet 50 SPP architecture.

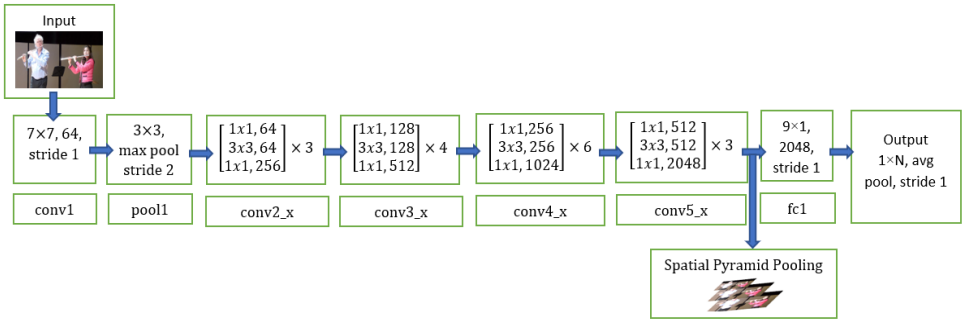


Fig. 3. Resnet 50 SPP architecture

The SPP blocks layer is included in the Resnet 50 configuration file as a result of our work. We also utilize the same SPP blocks layer in the configuration file with the spatial model, which is a nice touch. The spatial model employs down sampling in convolutional layers to get the relevant characteristics in the max-pooling layers, which are then used in the spatial model [31]. It applies three different sizes of the max pool for each image by using [route]. Different layers –2, –4 and –1, –3, –5, –6 in (conv)_5 have been used in each [route].

3. Results and discussions

3.1. Dataset

Photographs of individuals interacting with a range of musical instruments are included in the PPMI dataset. The datasets contain instruments such as the bassoon, cello, clarinet, French horn, erhu, flute, guitar, harp, saxophone, trumpet, recorder, and violin, among other things. Images of bassoon, erhu, flute, French horn, guitar, saxophone, and violin have been collected and published by Y a o and F e i-F e i [10].

Table 1. Musical instrument dataset

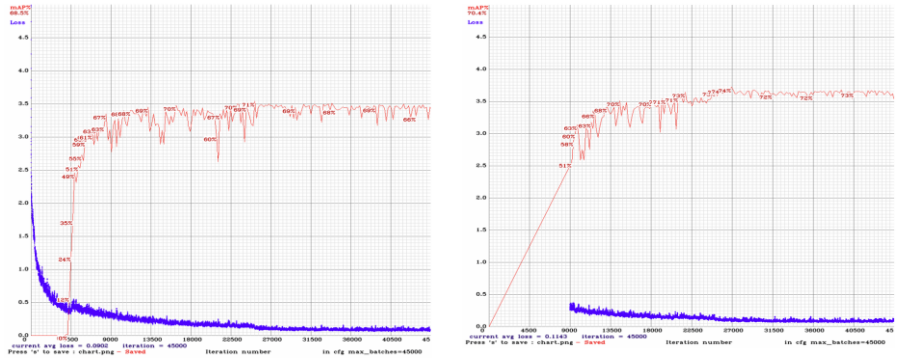
Class name	Training	Testing	Total image
Bassoon	253	109	362
Cello	225	97	322
Clarinet	221	95	316
Erhu	236	101	337
Flute	221	95	315
French horn	229	98	327
Guitar	228	98	326
Harp	232	100	332
Recorder	216	93	309
Saxophone	228	98	326
Trumpet	231	99	330
Violin	238	102	340
Total image	2759	1183	3942

A collection of photographs of instruments, including the cello, clarinet, harp, recorder and trumpet has been compiled by Aditya Khosla, and it has been released in September 2010 by Aditya Khosla. For training and testing purposes, the dataset initially had 100 photographs in each category for training and 100 images for testing

purposes. Table 1 provides a brief overview of the dataset. In this post, we used the PPMI dataset to train and evaluate the models we have developed. The collection includes photographs of people playing musical instruments taken from a range of perspectives, stances, and settings. Musical instrument performance style determines the variety of persons who participate in musical instrument performance. We increase the size of the dataset for each category by using data augmentation techniques such as rotation and flipping. The collection contains between 309 and 362 photos for each category. The total number of photos in our dataset increased to 3942, with 2759 images used for training and 1183 images used for testing.

3.2. Training result

The environment used to train the musical instrument recognition model consists of an Nvidia GTX2070 Super GPU accelerator, an AMD Ryzen 7 3700X Central Processing Unit (CPU) with an 8-core processor, and 32GB DDR4-3200 memory. This research work improves the Resnet 50 and Resnet 50 SPP models during the training stage by utilizing a learning rate of 0.001 for analysis, a learning rate decay of 0.1 at each epoch, and a momentum learning rate of 0.9. Therefore, Fig. 4a shows the consistency of the training process with Resnet 50. The training stage stops at 45,000 epochs. Resnet 50 applies max_batches = 45,000, mask_scale = 1, and the training loss value reach 0.0902. Furthermore, Resnet 50 SPP max_batches = 45,000, and mask_scale = 1. The iteration is unstable and experiences ups and downs ending at 45,000 epochs with a loss value of 0.1143 in Fig. 4b.



(a) (b)

Fig. 4. Training result using: Resnet 50 (a); Resnet 50 SPP (b)

Furthermore, the results of training performance are provided in Table 2. The training performance value for all classes, which includes the loss value, mAP, AP, precision, recall, F1, and IoU performance, as well as the overall performance value. Resnet 50 achieve the loss value 0.0902 with 30.8% IoU and 64.2% mAP. In other hand, Resnet 50 SPP exhibit 67.37% mAP with 43.51% IoU. This study uses IoU to determine the extent to which our projected border overlaps with the ground truth, which is the boundary of the actual object. Thus, our Resnet 50 SPP training model detected the objects with high accuracy. IoU calculates the overlap ratio between the boundary box of the prediction (pred), ground-truth (gt), and shown in the next equation [32]:

(1) $$\text{IoU} = \frac{\text{Area}_{\text{pred}} \cap \text{Area}_{\text{gt}}}{\text{Area}_{\text{pred}} \cup \text{Area}_{\text{gt}}}.$$

On the other hand, the output samples may be divided into three groups. True Positive (TP) is the number of samples that have been correctly identified; False Positive (FP) is the number of samples that have been wrongly recognized [33]; True Negative (TN) is the number of samples that have been incorrectly recognized, and False Negative (FN) is the number of samples that have been wrongly recognized. Precision and recall are represented by [34] in the equations

(2) $$\text{Precision } (P) = \frac{\text{TP}}{\text{TP+FP}},$$

(3) $$\text{Recall } (R) = \frac{\text{TP}}{\text{TP+FN}}.$$

Another evaluation index, F1 [35], is shown in the equation

(4) $$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+ Recall}}.$$

Yolo loss function based on the equation [14]

(5) $$\lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{\text{obj}} [(x_i - \hat{x}_i)^2 + (\text{y} - \hat{y}_i)^2] +$$
$$+\lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2 \right] +$$
$$+ \sum_{i=0}^{s^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i\right)^2 + \lambda_{\text{noobj}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i\right)^2 +$$
$$+ \sum_{i=0}^{s^2} \mathbb{I}_{i}^{\text{obj}} \sum_{c \in \text{classes}} (p_i(\text{c}) - \hat{p}_i(c))^2.$$

Table 2. Training performance results

Model	Loss value	Class ID	Name	AP (%)	TP	FP	Precision	Recall	F1-score	IoU (%)	**mAP@0.50** (%)
Resnet 50	0.0902	0	Bassoon	69.46	56	79	0.43	0.69	0.53	30.8	64.2
		1	Cello	68.16	53	105					
		2	Clarinet	46.66	47	121					
		3	Erhu	69.92	59	53					
		4	Flute	61.48	68	48					
		5	French horn	72.96	72	85					
		6	Guitar	76.95	59	84					
		7	Harp	83.59	60	142					
		8	Recorder	36.58	41	43					
		9	Saxophone	76.67	68	64					
		10	Trumpet	50.63	51	61					
		11	Violin	57.3	69	66					
Resnet 50 SPP	0.1143	0	Bassoon	78.93	55	15	0.6	0.63	0.61	43.51	67.37
		1	Cello	77.58	62	25					
		2	Clarinet	59.38	47	63					
		3	Erhu	68.37	49	25					
		4	Flute	63.13	49	21					
		5	French horn	71.02	65	48					
		6	Guitar	76.68	55	41					
		7	Harp	88.12	60	54					
		8	Recorder	42.78	47	43					
		9	Saxophone	80.68	64	25					
		10	Trumpet	49.58	39	33					
		11	Violin	52.24	51	30					

3.3. Discussion

The environment used to train the musical instrument recognition model consisted of an Nvidia GTX2070 Super GPU accelerator, an AMD Ryzen 7 3700X Central Processing Unit (CPU) with an 8-core processor, and 32GB DDR4-3200 memory.
Table 3 shows the results of testing accuracy result performance for 12 classes of musical instruments. Overall, Resnet 50 SPP is more precise than the previous version. Resnet 50 SPP increases the accuracy of previous method in all class. Moreover, Harp and Saxophone leading the highest accuracy 98% and 93% for Resnet 50 SPP. Followed by Clarinet 89%, Basson 85%, Recorder 85%, and Trumpet 85%. Saxophone and Guitar exhibit the highest accuracy 85% and 84% by utilizing Resnet 50. The optimum total average accuracy obtained by Resnet 50 SPP with 84.64% accuracy and 35.93 millisecond for detection time. However, Resnet 50 gain 74.92% average accuracy with the average of detection time is 30.21 milliseconds. Besides, Table 3 describes the comparison results between Resnet 50 and Resnet 50 SPP in terms of detection time. Based on these results, it can be concluded that Resnet 50 is faster than Resnet 50 SPP in terms of detection time.

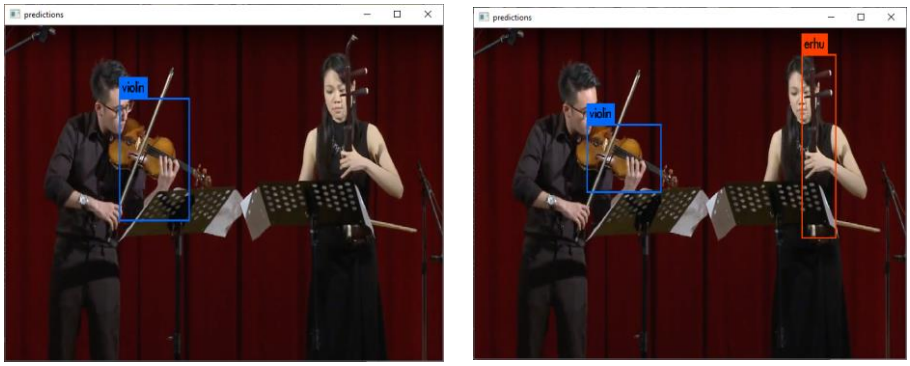
Table 3. Testing accuracy results performance

Class	Resnet 50		Resnet 50 SPP	
	Accuracy (%)	Time (ms)	Accuracy (%)	Time (ms)
Bassoon	79%	33.46	85%	38.32
Cello	73%	33.12	81%	38.06
Clarinet	76%	29.39	89%	38.03
Erhu	72%	29.36	81%	38.03
Flute	81%	29.50	82%	38.00
French horn	74%	29.40	78%	37.65
Guitar	84%	29.30	79%	35.81
Harp	83%	29.40	98%	34.65
Recorder	65%	29.80	85%	33.15
Saxophone	85%	29.70	93%	33.08
Trumpet	65%	30.40	85%	33.16
Violin	62%	29.70	80%	33.20
Average	74.92%	30.21	84.64%	35.93

Clarinet and flute are extremely similar musical instruments in terms of their form, the manner they are played, and the size of the instruments. Guitar, violin, and cello are all musical instruments that are comparable to one another. Those three musical instruments are similar in colour, form, and size, yet they are significantly different in size. The smallest instrument is the violin, the middle-sized instrument is the guitar, and the largest instrument is the cello.

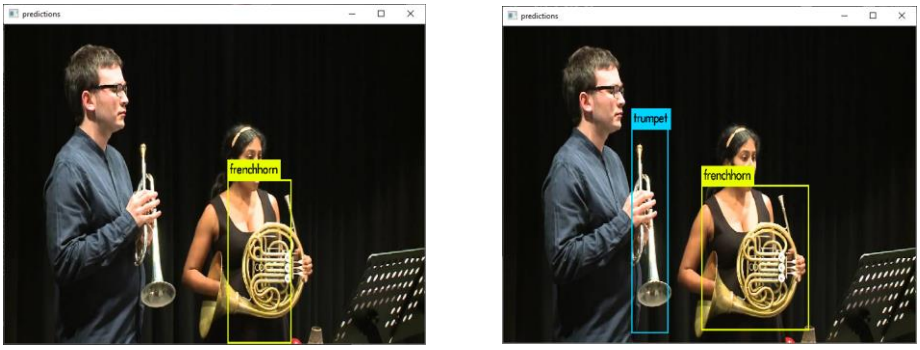
Fig. 5 shows the recognition result of violin and erhu. Resnet 50 allocates additional workspace size of 26.22 MB and loads 69 layers from weights-file. Image a1.jpg is predicted in 33.85 milliseconds with the result violin obtains 67% accuracy shown in Fig. 5a. Resnet 50 can detect only one music instrument in the image. Furthermore, recognition result of Resnet 50 SPP by using the same image is described in Fig. 5b. Image a1.jpg is predicted in 38.8 milliseconds as a result violin obtains 81%, and erhu 57% accuracy. The Resnet 50 SPP achieves the highest

average accuracy for all violin and guitar classes but takes longer time to detect objects in the image.



(a) Resnet 50 (b) Resnet 50 SPP

Fig. 5. Violin and Erhu recognition results

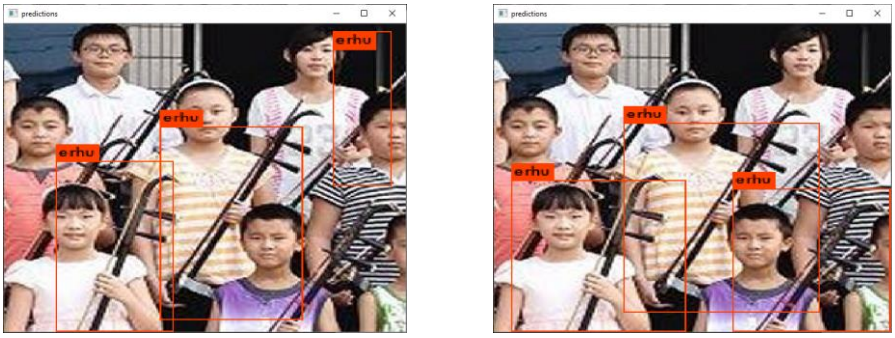


(a) Resnet 50 (b) Resnet 50 SPP

Fig. 6. Trumpet and French horn recognition result

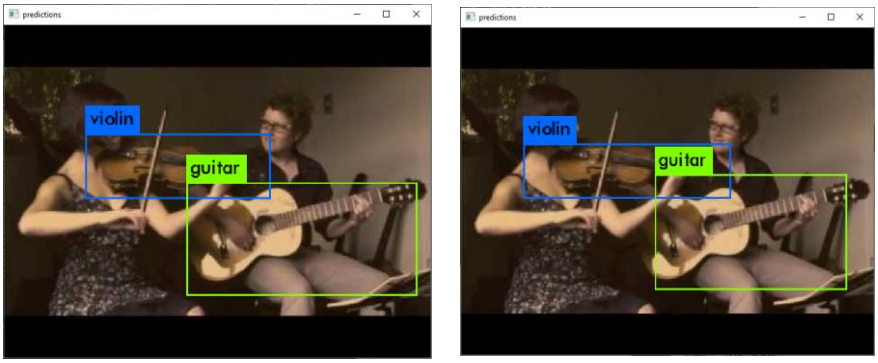
Fig. 6a illustrates the result of trumpet and French horn recognition using Resnet 50. After loading 69 layers from the weights-file with a total BFLOPS of 26.453, the expected time for Images a2.jpg is 38.2 ms. As a consequence, French horn accuracy is 83 %. Resnet 50 failed to detect the French horn, it only detected one instrument. In comparison, Resnet 50 SPP projected trumpet attains 79 % and French horn attain 90 % in 38.9 ms, as seen in Fig. 6b.

The recognition result of erhu with multiple objects can be seen in Fig. 7. The minimum accuracy is obtained by Resnet 50 in Fig. 7a and image *violin1.jpg* has been predicted in 33.711 ms. Resnet 50 has recognized 3 recorders in the image with the accuracy 89%, 90%, and 30%, successively. As demonstrated in Fig. 7b, the *violin1.jpg* picture has been predicted in 38.713 ms employing Resnet 50 SPP and it is capable of recognizing three erhus with accuracy of 81 %, 92 %, and 73%, accordingly. Fig. 8a illustrate the Violin and Guitar recognition result using Resnet 50. Image *b1.jpg* has predicted in 35.861 ms with the accuracy violin 55% and guitar 90%. The optimum accuracy has been achieved by Resnet 50 SPP and is described in Fig. 8b. Violin attains 74% of accuracy and guitar achieves 98% of accuracy.



(a) Resnet 50 (b) Resnet 50 SPP

Fig. 7. Erhu recognition result



(a) Resnet 50 (b) Resnet 50 SPP

Fig. 8. Violin and Guitar recognition result

Based on the experimental results, it can be concluded that Resnet 50 SPP has better performance compared to Resnet 50. Moreover, Resnet 50 SPP can detect all musical instruments in the image, while Resnet 50 fails to detect musical instruments in Fig. 5 and Fig. 6. Hence, Resnet 50 SPP achieves the highest average accuracy for all classes but takes longer time to detect objects in the image.

4. Conclusion

The major focus of this research is on how we try to distinguish between items that are highly similar at the human eye level. Our research investigations make use of Resnet 50 in conjunction with Spatial Pyramid Pooling (SPP) to identify musical instruments that are comparable to one another in terms of their visual appearance. In this study, we detect numerous musical instruments that are comparable to one other, such as the bassoon, cello, clarinet, erhu, flute, French horn, harp, recorder, saxophone, trumpet, and violin. Our research explores and evaluates CNN models paired with a variety of backbone architectures and extractor features, most notably Resnet 50 for object identification. This experiment investigates the detector's primary characteristics, such as precision accuracy, detection time, workspace size, and BFLOP number. Based on our experimental result we can improve the performance of recognising similar objects, e.g., music instruments. Our proposed

method Resnet 50 SPP exhibit the highest average accuracy of 84.64% compared to the results of previous studies. Resnet 50 SPP enhances the detecting process and beats other approaches. We intend further research for recognition of inaccurately formed musical instruments in a picture as a part of our future study. Additionally, we intend to include Explainable Artificial Intelligence (XAI) into our future study to give additional insight on the picture.

***Acknowledgment*:** This paper has been supported by the Ministry of Science and Technology, Taiwan. The Nos are MOST-110-2927-I-324-50, MOST-110-2221-E-324-010, MOST-109-2622-E-324-004, Taiwan.

References

1. Ribeiro, A. C. M., R. C. Scharlach, M. M. C. Pinheiro. Assessment of Temporal Aspects in Popular Singers. – CODAS, Vol. **27**, 2015.
https://doi.org/10.1590/2317-1782/20152014234
2. Bai, T., Y. Pang, J. Wang, K. Han, J. Luo, H. Wang, J. Lin, J. Wu, H. Zhang. An Optimized Faster R-CNN Method Based on DRNet and RoI Align for Building Detection in Remote Sensing Images. – Remote Sens., Vol. **12**, 2020.
https://doi.org/10.3390/rs12050762
3. Wetzel, J., A. Laubenheimer, M. Heizmann. Joint Probabilistic People Detection in Overlapping Depth Images. – IEEE Access, Vol. **8**, 2020.
https://doi.org/10.1109/ACCESS.2020.2972055
4. Dewi, C., R. C. Chen, H. Yu. Weight Analysis for Various Prohibitory Sign Detection and Recognition Using Deep Learning. Multimed. – Tools Appl. Vol. **79**, 2020, pp. 32897-32915.
https://doi.org/10.1007/s11042-020-09509-x
5. Xi, X., Z. Yu, Z. Zhan, Y. Yin, C. Tian. Multi-Task Cost-Sensitive-Convolutional Neural Network for Car Detection. – IEEE Access, Vol. **7**, 2019.
https://doi.org/10.1109/ACCESS.2019.2927866
6. Dewi, C., R. C. Chen, Y. T. Liu. Wasserstein Generative Adversarial Networks for Realistic Traffic Sign Image Generation. – In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021, pp. 479-493.
https://doi.org/10.1007/978-3-030-73280-6_38
7. Ju, M., S. Moon, C. D. Yoo. Object Detection for Similar Appearance Objects Based on Entropy. – In: Proc. of 7th International Conference on Robot Intelligence Technology and Applications (RiTA'19), 2019.
https://doi.org/10.1109/RITAPP.2019.8932791
8. Jiang, Y., L. Chen, H. Zhang, X. Xiao. Breast Cancer Histopathological Image Classification Using Convolutional Neural Networks with Small SE-ResNet Module. – PLoS One, Vol. **14**, 2019.
https://doi.org/10.1371/journal.pone.0214587
9. Yu, X., C. Kang, D. S. Guttery, S. Kadry, Y. Chen, Y. D. Zhang. ResNet-SCDA-50 for Breast Abnormality Classification. IEEE/ACM Trans. – Comput. Biol. Bioinforma, Vol. **18**, 2021.
https://doi.org/10.1109/TCBB.2020.2986544
10. Yao, B., L. Fei-Fei. Grouplet: A Structured Image Representation for Recognizing Human and Object Interactions. – In: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010.
https://doi.org/10.1109/CVPR.2010.5540234
11. Zhang, X., F. Wan, C. Liu, X. Ji, Q. Ye. Learning to Match Anchors for Visual Object Detection. – IEEE Trans. Pattern Anal. Mach. Intell., 2021.
https://doi.org/10.1109/TPAMI.2021.3050494

12. G i r s h i c k, R. Fast R-CNN. – In: Proc. of IEEE International Conference on Computer Vision, 2015, pp. 1440-1448.
https://doi.org/10.1109/ICCV.2015.169
13. C h e n g, G., Y. S i, H. H o n g, X. Y a o, L. G u o. Cross-Scale Feature Fusion for Object Detection in Optical Remote Sensing Images. – IEEE Geosci. Remote Sens. Lett., Vol. **18**, 2021.
https://doi.org/10.1109/LGRS.2020.2975541
14. R e d m o n, J., S. D i v v a l a, R. G i r s h i c k, A. F a r h a d i. You Only Look Once: Unified, Real-Time Object Detection. – In: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, pp. 779-788.
https://doi.org/10.1109/CVPR.2016.91
15. L i u, W., D. A n g u e l o v, D. E r h a n, C. S z e g e d y, S. R e e d, C. Y. F u, A. C. B e r g. SSD: Single Shot Multibox Detector. – In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, pp. 21-37.
https://doi.org/10.1007/978-3-319-46448-0_2
16. S r i n i v a s a n, K., P. B a l a m u r u g a n, V. R. A z h a g u r a m y a a. Survey on Similar Object Detection in H.264 Compressed Video. – In: Proc. of 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET'17), 2017.
https://doi.org/10.1109/ICAMMAET.2017.8186663
17. G r a u m a n, K., T. D a r r e l l. The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features. – In: Proc. of IEEE International Conference on Computer Vision, 2005, pp. 1458-1465.
https://doi.org/10.1109/ICCV.2005.239
18. L a z e b n i k, S., C. S c h m i d, J. P o n c e. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. – In: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 1-8.
https://doi.org/10.1109/CVPR.2006.68
19. D a i, J., Y. L i, K. H e, J. S u n. R-FCN: Object Detection via Region-Based Fully Convolutional Networks. – In: Advances in Neural Information Processing Systems, 2016, pp. 379-387.
20. S i v i c, J., A. Z i s s e r m a n. Video Google: A Text Retrieval Approach to Object Matching in Videos. – In: Proc. of IEEE International Conference on Computer Vision, 2003, pp. 1-8.
https://doi.org/10.1109/iccv.2003.1238663
21. Y a n g, J., K. Y u, Y. G o n g, T. H u a n g. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. – In: Proc. of 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, 2009, pp. 1794-1801.
https://doi.org/10.1109/CVPRW.2009.5206757
22. W a n g, J., J. Y a n g, K. Y u, F. L v, T. H u a n g, Y. G o n g. Locality-Constrained Linear Coding for Image Classification. – In: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3360-3367.
https://doi.org/10.1109/CVPR.2010.5540018
23. V a n d e S a n d e, K. E. A., J. R. R. U i j l i n g s, T. G e v e r s, A. W. M. S m e u l d e r s. Segmentation as Selective Search for Object Recognition. – In: Proc. of IEEE International Conference on Computer Vision, 2011, pp. 1879-1886.
https://doi.org/10.1109/ICCV.2011.6126456
24. H e, K., X. Z h a n g, S. R e n, J. S u n. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. – IEEE Trans. Pattern Anal. Mach. Intell., Vol. **37**, 2015, pp. 1904-1916.
https://doi.org/10.1109/TPAMI.2015.2389824
25. H e, K., X. Z h a n g, S. R e n, J. S u n. Deep Residual Learning for Image Recognition. – In: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.
https://doi.org/10.1109/CVPR.2016.90
26. C h a n d e r, G., B. L. M a r k h a m, D. L. H e l d e r. Summary of Current Radiometric Calibration Coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI Sensors. – Remote Sens. Environ., Vol. **113**, 2009, pp. 893-903.
https://doi.org/10.1016/j.rse.2009.01.007

27. F a n g, W., C. W a n g, X. C h e n, W. W a n, H. L i, S. Z h u, Y. F a n g, B. L i u, Y. H o n g. Recognizing Global Reservoirs from Landsat 8 Images: A Deep Learning Approach. – IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., Vol. **12**, 2019, pp. 3168-3177.
https://doi.org/10.1109/jstars.2019.2929601
28. I b r a h i m, Y., H. W a n g, M. B a i, Z. L i u, J. W a n g, Z. Y a n g, Z. C h e n. Soft Error Resilience of Deep Residual Networks for Object Recognition. – IEEE Access, Vol. **8**, 2020, pp. 19490-19503.
https://doi.org/10.1109/ACCESS.2020.2968129
29. W e n, L., X. L i, L. G a o. A Transfer Convolutional Neural Network for Fault Diagnosis Based on ResNet-50. – Neural Comput. Appl., Vol. **32**, 2020.
https://doi.org/10.1007/s00521-019-04097-w
30. F u l t o n, L. V., D. D o l e z e l, J. H a r r o p, Y. Y a n, C. P. F u l t o n. Classification of Alzheimer's Disease with and without Imagery Using Gradient Boosted Machines and Resnet-50. – Brain Sci., Vol. **9**, 2019.
https://doi.org/10.3390/brainsci9090212
31. D e w i, C., R.-C. C h e n, Y.-T. L i u, S.-K. T a i. Synthetic Data Generation Using DCGAN for Improved Traffic Sign Recognition. – Neural Comput. Appl., Vol. **33**, 2021, pp. 1-15.
32. A r c o s-G a r c í a, Á., J. A. Á l v a r e z-G a r c í a, L. M. S o r i a-M o r i l l o. Evaluation of Deep Neural Networks for Traffic Sign Detection Systems. – Neurocomputing., Vol. **316**, 2018, pp. 332-344.
https://doi.org/10.1016/j.neucom.2018.08.009
33. D e w i, C., R. C. C h e n, H. Y u, X. J i a n g. Robust Detection Method for Improving Small Traffic Sign Recognition Based on Spatial Pyramid Pooling. – J. Ambient Intell. Humaniz. Comput., Vol. **12**, 2021.
https://doi.org/10.1007/s12652-021-03584-0
34. Y a n g, H., L. C h e n, M. C h e n, Z. M a, F. D e n g, M. L i, X. L i. Tender Tea Shoots Recognition and Positioning for Picking Robot Using Improved YOLO-V3 Model. – IEEE Access., Vol. **7**, 2019, pp. 180998-181011.
https://doi.org/10.1109/ACCESS.2019.2958614
35. T i a n, Y., G. Y a n g, Z. W a n g, H. W a n g, E. L i, Z. L i a n g. Apple Detection During Different Growth Stages in Orchards Using the Improved YOLO-V3 Model. – Comput. Electron. Agric., Vol. **157**, 2019, pp. 417-426.
https://doi.org/10.1016/j.compag.2019.01.012

Received: 16.11.2021; Second Version: 15.02.2022; Accepted: 25.02.2022