Integrating shape- and CNN-based features for zero-shot and low-shot learning

Sandipani Basu[1,†], Noyon Dey[1,†], Suchendra M Bhandarkar[1,*,†] and Steven Wolbach[2]

[1] School of Computing, University of Georgia, Athens, GA 30602, USA

[2] U.S. Army Combat Capabilities Development Command (DEVCOM) Analysis Center, Aberdeen Proving Ground, MD 21005, USA

Abstract

Image classification involves categorizing images into predefined classes based on their visual features. While most existing classifiers perform well on noise-free and non-corrupted images, their performance is significantly compromised on real-world images that exhibit severe degradation. Real-world images, especially ones of military targets, are typically captured under extenuating battlefield conditions and subject to a variety of natural and anthropogenic stressors such as occlusion, obfuscation, camouflage, distortion and sensor noise. Traditional image classifiers prove inadequate in such situations since they require large amounts of visual training data to build resilience to the above stressors. Since visual training data acquired under a variety of viewing conditions are not always abundant, the problem of achieving acceptable classification accuracy with limited training data is of growing interest. Zero-shot learning (ZSL) and low-shot learning (LSL) offer a means for incorporating auxiliary information sources to compensate for the absence or paucity of visual training data respectively. We propose a supervised LSL-based classifier termed as *CoNNText* that uses auxiliary shape-based information to improve the robustness of traditional convolutional neural network (CNN)-based RGB image classifiers. The proposed CoNNText model integrates the shape context descriptor with CNN-derived RGB image features to yield improved automatic target recognition (ATR) accuracy for military vehicles in RGB images corrupted by various environmental and anthropogenic stressors with limited visual training data. Experimental results show that the CoNNText model improves upon the benchmark CNN classification accuracy, quantified using the F-1 score and AUC (area under the ROC curve) as the performance metrics, when tested on an RGB image dataset of military vehicles under varying battlefield stressors.

Keywords

Image Classification, Low-Shot Learning, Shape Context Descriptor, Feature Fusion, Convolutional Neural Network

1. Introduction

Automatic target recognition (ATR) is an important component of autonomous military systems. The development and deployment of accurate real-time ATR systems is of paramount importance to the military given their critical role in modern warfare. In recent times, ATR systems have undergone a significant revolution with the incorporation of artificial intelligence (AI) and

CVCS2024: the 12th Colour and Visual Computing Symposium, September 5–6, 2024, Gjøvik, Norway

[*] Corresponding author.

[†] These authors contributed equally.

sandipani.basu@uga.edu (S. Basu); noyon.dey@uga.edu (N. Dey); suchi@uga.edu (S. M. Bhandarkar); steven.d.wolbach.civ@army.mil (S. Wolbach)

009-0000-5895-5991 (S. Basu); 0000-0001-5074-3281 (N. Dey); 0000-0003-2930-4190 (S. M. Bhandarkar)

machine learning (ML) techniques. The integration of AI and ML techniques in military systems and operations has transformed various aspects of modern warfare resulting in the extensive deployment of a variety of systems with semi and completely autonomous capabilities including unmanned aerial vehicles (UAVs), unmanned ground vehicles (UGVs), autonomous decision-support systems, ATR systems, autonomous target tracking and engagement capabilities among others [17].

Enabling high levels of autonomy and accuracy in ATR systems under battlefield conditions requires enormous amounts of training data acquired under a wide variety of battlefield conditions. Due to the adverse nature of battlefield conditions characterized by the presence of several detrimental factors and a wide variety of viewing conditions, accurate target identification and localization requires that the classifier be trained on data that encompass a large range of detrimental factors and viewing conditions. The paucity of such training data presents a significant bottleneck in the development and deployment of robust and reliable ATR systems. Consequently, ATR systems need to be designed to be capable of identifying target classes that they have not been explicitly or adequately trained on. These include previously unseen or rarely seen target classes and more importantly, known targets that are corrupted with degradation factors (e.g., noise, occlusion, illumination variations, camouflage, distortion) that are not represented in the training image dataset.

The aforementioned paucity of training data in ATR is addressed by Zero-Shot Learning (ZSL) and Low Shot Learning (LSL) approaches. ZSL is essentially designed for situations where a target class in the testing data is entirely absent from the training data whereas in LSL the target class in the testing data is insufficiently represented in the training data [2], [3]. LSL, also referred to as few-shot learning (FSL), deals with situations where the final classification (during testing time) must be done based on a few observed training samples whereas in the case of ZSL, the final classification is performed for a previously unobserved target class. ZSL and LSL methods represent a generalization of the traditional supervised learning methods to improve classification accuracy in that LSL encourages the model to be robust and invariant to environmental variations in visual features whereas ZSL goes a step further, wherein the model learns to classify novel target classes by transferring the relevant knowledge from previously observed target classes.

Humans perform ZSL and LSL naturally. For example, a person who has never seen a zebra before would be able to recognize it in an image if the person were told that a zebra is visually similar to a horse (a previously observed class) with the distinguishing characteristic of possessing stripes (provided as auxiliary information). To achieve ZSL and LSL capabilities within an ML system, we need to identify the sources of auxiliary knowledge that would allow generalization from previously seen target classes to unseen or rarely seen target classes. To this end we propose a fusion model termed *CoNNText*, that integrates auxiliary information with color (RGB) image data to enable this generalization. We use convolutional neural network (CNN)-based models for extraction of RGB image features and fuse them with shape-based information in the form of the *shape context* (SC) descriptor to enable ZSL/LSL-based target classification [4], [5]. The proposed CoNNText model enables us to determine the functional dependence of the classification accuracy on the global CNN-based image features and a geometric shape representation in the form of the SC descriptor. We expect the CoNNText model to yield higher classification accuracy compared to traditional CNN-based models when the input images of seen target classes are subject to various battlefield stressors (e.g., distortion, camouflage, defilade, vertical coloration, and horizontal coloration) with limited training data.

The remainder of the paper is organized as follows: Section 2 presents a review of the relevant ATR research literature; the datasets used are described in Section 3; the relevant ATR methods are described in Section 4; the experimental results are presented and discussed in Section 5; and the paper is concluded in Section 6 with an outline for future work.

2. Literature Review

Convolutional Neural Networks (CNNs) have radically changed the field of computer vision on account of their ability to learn hierarchical representations directly from raw image data, thereby eliminating the need for feature engineering. The pioneering CNN architecture LeNet-5, introduced by LeCun et. al [3], demonstrated the power of convolutional, pooling and fully-connected layers in the context of handwritten digit recognition. The AlexNet proposed by Krizhevsky et al. [4], is a deep CNN architecture that won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. AlexNet with a deeper architecture comprising of multiple convolutional and fully connected layers, dropout regularization and the use of Rectified Linear Units (ReLUs) [24] significantly outperformed other contemporary CNN models. AlexNet paved the way for several deep CNN architectures, namely VGGNet [5] whose primary feature was the increased depth of the network architecture. The ResNet [6], a deep CNN architecture with residual connections to address the vanishing gradient problem, achieved state-of-the-art performance on several visual recognition tasks and became a fundamental building block for deeper architectures. The use of CNNs was extended beyond image classification. With the advent of region-based CNNs (RCNNs) [7] that combined CNN features and region proposal algorithms, object detection in images became feasible and led to the development of Faster-RCNN [8] and Mask-RCNN [9] architectures which achieved state-of-the-art performance in object detection, semantic segmentation and instance segmentation tasks. Recent advancements in CNNs include the inclusion of attention mechanisms [18], which enable the CNN model to focus on relevant image regions, and self-supervised learning, which leverages unsupervised pretraining to boost the CNN performance with limited labeled data.

Zero-shot learning (ZSL) is an ML paradigm where a pretrained deep learning model is made to generalize on a novel category of samples. The seminal work by Lampert et al. [10] introduced the concept of attribute-based ZSL, where each class is associated with a set of semantic attributes. Akata et al. [11] introduced the use of CNNs to map visual features to semantic embeddings, whereas Socher et al. [12] utilized recursive neural networks (RNNs) to map visual features to compositional phrase embeddings. The semantic attributes and embeddings were used to bridge the gap between seen and unseen classes within the ZSL framework. Low-shot learning (LSL) is an ML framework that enables a pretrained model to generalize over new categories of data using only a few labeled samples per class, and is regarded as a form of meta-learning [27]. Vinyals et al. [13] proposed *Matching Networks* to learn a metric space in which instances from the same class are mapped in closer proximity than instances from different classes. To further enhance LSL performance, meta-learning approaches such as Model-Agnostic Meta-Learning (MAML) [14] and Meta-Learning with Memory-Augmented

Neural Networks (MANNs) have been proposed. Recent advancements have also explored the combination of ZSL and LSL techniques, termed Generalized Zero-shot Learning (GZSL) [19]. GZSL aims to bridge the gap between seen and unseen classes while addressing the challenges of limited labeled data, enabling more comprehensive and flexible learning scenarios.

The shape context (SC) descriptor, introduced by Belongie et al. [1], [2], captures the distribution of local shape features around each point on a shape contour. A formal definition of the SC descriptor is as follows: Consider n points sampled on the shape contour. Consider the set of vectors originating from a sample point p_i on the shape contour to the remainder $n-1$ contour sample points on the shape. For the point p_i, a coarse histogram h_i representing the distribution of the relative positions of the remaining $n-1$ sample contour points is computed and defined as the shape context of the point p_i [1]. The key idea behind the SC descriptor is to represent each contour point by a histogram that encodes the relative spatial relationship between the reference sample point and the remaining sample points on the shape contour. The histogram bins represent different angular sectors and log-radial distances, thus capturing the local shape structure in $(\log \rho, \theta)$ space where ρ is the radial distance and θ is the polar angle.

Multimodal feature fusion has gained significant attention in recent years with the increasing availability of data from various modalities, such as images, text, audio, and sensor data [25]. By exploiting the complementary and synergistic nature of multiple input data modalities, the capabilities of the ML models are enhanced using multimodal feature fusion. Early approaches in multimodal feature fusion focused on early-stage fusion where feature vectors from different modalities are simply concatenated at the input level [25]. However, early-stage fusion methods often face challenges that arise from the heterogeneity and varying dimensionalities of the underlying multimodal feature spaces which limit the effectiveness of feature fusion. Late-stage fusion techniques [26] have emerged as an alternative, where features from individual modalities are processed separately via individual ML models whose predictions or representations are then combined at a later stage. This approach allows for flexibility in modeling and can handle feature spaces of varying dimensionalities and complexities. Techniques such as decision-level fusion, score-level fusion, and feature-level fusion have been explored in the late-stage fusion paradigm [26]. Applications of multimodal feature fusion span various domains, including multimedia analysis, human-computer interaction, healthcare, and autonomous systems. Multimodal feature fusion has been successfully applied in various tasks such as multimodal sentiment analysis, audio-visual speech recognition, multimedia retrieval, and multimodal medical image analysis [26].

3. Description of Datasets

In this work, we developed our in-house dataset consisting of $\approx$ **26,000** RGB images collected from public internet sources. An 80-10-10 split among the training, validation and testing sets was used. Per the requirements of the U.S. Army, these images were of five different army vehicles namely; **Abrams** (tank), **Bradley** (tracked armored fighting vehicle), **MRAP** (mine-resistant ambush-protected light tactical vehicle), **HMMWW** (high mobility multipurpose wheeled vehicle), and **Stryker** (armored infantry personnel carrier) as depicted in Fig. 1.

One of the primary aims of our work is to test the proposed ZSL/LSL framework on stressed

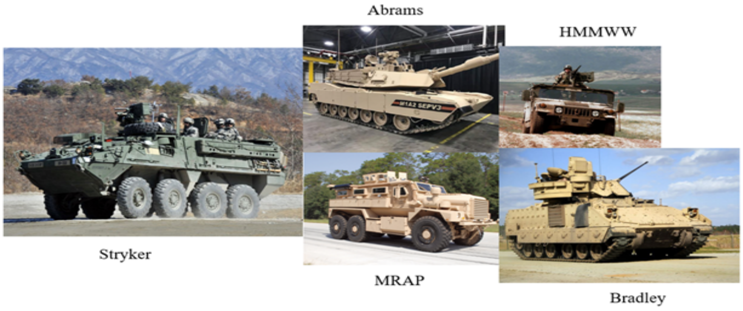

Figure 1: Sample military vehicle images obtained from the web.

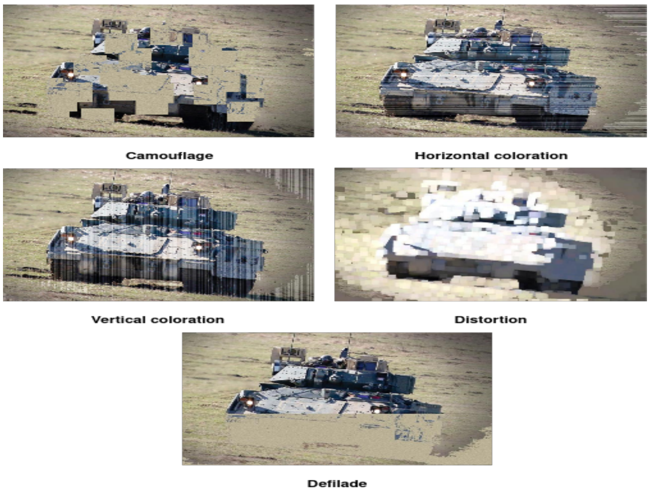

Figure 2: Images of the Bradley vehicles at 50% stress level for a variety of stressors.

images obtained from a dataset comprising of RGB images of military vehicles. Using software packages developed by the US Army DEVCOM Analysis Center [15], we simulated some common battlefield stressors on these images to test the robustness of our ZSL/LSL classification models. All the stressors were applied over the target military vehicle image in 5% increments of the image area [15]. The corrupted images resulting from the application of various stressors were used to test the robustness of the proposed ZSL/LSL framework to the corresponding stressed environmental conditions. The stressors chosen were those commonly encountered in battlefield environments such as *defilade*, *coloration*, *distortion*, and *camouflage* as depicted in Fig. 2. Additionally, a dataset of 3D CAD model renderings of military vehicles was also created. The 3D CAD model renderings were obtained by viewing the 3D CAD model in 5^oincrements long the azimuth dimension, and 90^o and 60^o increments along the polar dimension [15]. We refer to this dataset as the 3D model dataset, and use it to train and test our SC descriptor [1].

4. Automatic Target Recognition Methods

In order to evaluate and compare different ZSL/LSL-based ATR methods on a test dataset of images corrupted with natural and anthropogenic (i.e., human-induced) stressors, e.g., camouflage, coloration, defilade, and distortion, we designed and implemented the following classification schemes: (a) Pure CNN-based classification and, (b) CNN and SC descriptor fusion-based classification.

4.1. Pure CNN-based Classification

We employed a state-of-the-art CNN that is pre-trained on the ImageNet[16] database and specifically designed to classify images into predefined classes. Given the limited size of our training set of images containing military vehicles and our unique classification domain, we use transfer learning to fine-tune the pretrained CNN models. Transfer learning [22], [23] results in significantly higher performance on a new task compared to training from scratch on a new task using a small training dataset. We use three well-known CNN-based image classification models: Resnet-50 [6], Inception-V3 [20] and VGG-16 [5][24]. In all of these CNN models, we replaced the final classification head with our own classification layer during transfer learning to enable the model to learn task- and target-specific features. Henceforth, we refer to the pure CNN-based classifier as *PureCNN*.

4.2. CoNNText: CNN and SC Descriptor Fusion-based Classification

A novel model termed *CoNNText* is proposed for ZSL/LSL-based ATR. CoNNText employs fusion of the SC descriptor with CNN-derived color and texture features to enhance the robustness of target classification in the presence of stressors. The fusion of the CNN-derived deep features and SC descriptor is expected to result in more robust classification, especially for images wherein the target objects are occluded and/or subject to natural and human-induced stressors.

4.2.1. Mask Generation:

Generation of the SC descriptor for an object is preceded by the computation of an accurate object instance mask in the image that essentially captures the object silhouette. We used the M-RCNN model [9] trained on 100 high-quality RGB images of target objects for this purpose. The criteria for selecting the high-quality RGB images are as follows: (a) the target object in the image should be free of occlusion, (b) the images should not contain secondary objects such as humans, trees, and other military vehicles, and (c) the target object should not be trenched or camouflaged in the image.

4.2.2. Feature Fusion:

The fusion of the CNN-derived deep image features and the SC descriptors is performed at an early stage of the processing pipeline (i.e., early-stage fusion). Since the input data modalities are fused before any significant high-level analysis or decision-making is performed, the early-stage fusion can preserve the raw information derived from multiple modalities, and allow the model

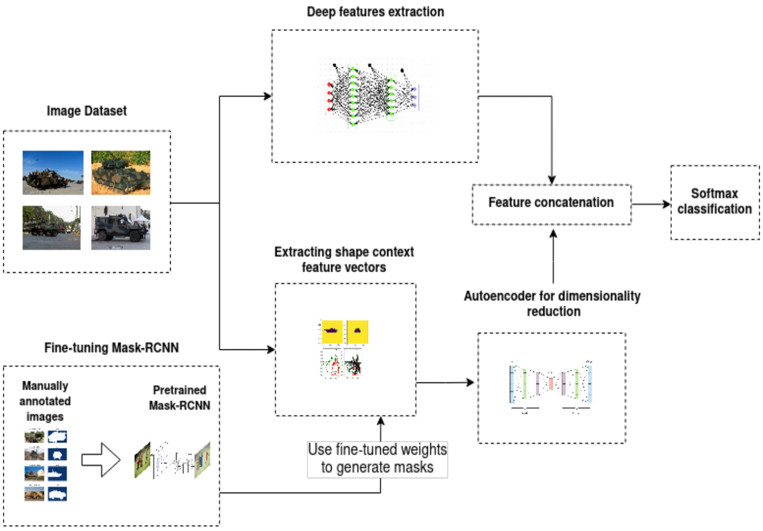
Deep features extraction
Image Dataset
Feature concatenation
Softmax classification
Extracting shape context feature vectors
Autoencoder for dimensionality reduction
Fine-tuning Mask-RCNN
Manually annotated images
Pretrained Mask-RCNN
Use fine-tuned weights to generate masks

Figure 3: Proposed CoNNText model architecture for ATR using feature fusion.

to learn jointly from different input sources by capturing both low-level details and high-level correlations between the multiple modalities.

The schematic of the proposed early-stage fusion is shown in Fig. 3. Since the features to be fused are both visual features, we concatenated the feature vectors obtained from the CNN and SC descriptors before feeding them into subsequent layers or classifiers. A major challenge faced in the early-stage fusion process stems from the heterogeneity and different dimensionalities of the underlying feature modalities since CNN-derived deep feature vectors are 512-dimensional whereas the SC descriptor is 6000-dimensional. Consequently, dimensionality reduction, feature selection and feature mapping techniques need to be employed to appropriately align the feature vectors derived from the two modalities. We employed an autoencoder [21] to map the two input feature vectors into a shared vector space of reduced dimensionality to preserve the most informative features from both modalities while discarding unnecessary and redundant information. We also experimented with Principle Component Analysis (PCA), Independent Component Analysis (ICA), and Linear Discriminate Analysis (LDA) techniques in this regard but found the autoencoder to yield the best results.

5. Experimental Results And Discussion

We performed a holistic comparative analysis of the effect of no stress, low stress, medium stress, and high stress on the performance of the PureCNN and CoNNText models. The PureCNN model used the Resnet-50, Inception-V3, or VGG-16 model for target classification. Likewise, the CoNNText model also used one of the aforementioned CNN models as the base model for early-stage fusion followed by target classification. Both models were fed RGB images corrupted with the presence of stressors. For each stressor type, the stress levels chosen were 0%, 25%, 50%, and 75% where 0% represents an uncorrupted (unstressed) input image. We computed and compared the classification accuracy of both the PureCNN model and CoNNText model for each stressor where the classification accuracy represented the proportion of correctly classified samples out of the total number of samples in the dataset (expressed as a percentage). In particular, we observed and analyzed how the fusion of CNN-derived image features and shape information impacted the overall classification accuracy. We also computed the F1-scores and AUC (area under the ROC curve) values across stressors and target classes. The results in Figs. 4, 5 and 6 are presented as an array of bar charts where each row represents a target class (Abrams, Bradley, HMMWW, MRAP or Stryker) and each column represents a stressor type (defilade, distortion, vertical coloration, horizontal coloration, or camouflage). Within each bar chart, the x-axis represents the stress levels (0% - 75%) and y-axis represents the performance metric (F-1 score or AUC). We observe how each of the three CoNNText models (with Resnet-50, Inception-V3, or VGG-16 as the base CNN) performs in comparison to one another in terms of the F-1 score (Fig. 4). We also observe how the CoNNText model performs in comparison to its corresponding PureCNN model in terms of the F-1 score and AUC metrics (Figs. 5 and 6).

Before presenting a per-stressor type and per-target class analysis of performance, we make the following general observations: (a) The CoNNText Resnet-50 is observed to clearly outperform its PureCNN counterpart with the inclusion of the SC descriptor yielding higher F1-scores across all the stressor types and stress levels. (b) The CoNNText Inception-V3 does not perform worse than its PureCNN counterpart, exceeding the PureCNN's F-1 score only for the Horizontal Coloration stressor. (c) The CoNNText VGG-16 however is not able to capture the shape features effectively to boost its F-1 score in comparison to its PureCNN counterpart.

5.1. Detailed Per Stressor Type Analysis

For the defilade stress, at 25% and 50% stress levels, CoNNText performs better than its PureCNN counterpart (Figs. 5 and 6). In the case of the MRAP class, CoNNText shows considerable improvement over PureCNN owing to MRAP's unique shape and structure. In the case of MRAP and Bradley, CoNNText shows consistently better performance over its PureCNN counterpart (Figs. 5 and 6). At the 75% stress level, CoNNText exhibits a drop in F-1 score for the MRAP class. Since both, HMMWW and MRAP have boxy shapes, CoNNText, which relies on the SC descriptor-based shape information, predicts some instances of MRAP as HMMWW. Although there is a distinct difference between the MRAP and HMMWW shapes (HMMWW has more straight edges and flat surfaces whereas MRAP has a more angular and curved shape) during a defilade stress test, most object parts are obscured resulting in the confusion between the two classes. Additional auxiliary information would be needed to resolve this confusion under defilade stress.

For the distortion stress, at 25% stress level, we observe that CoNNText performs almost the same as PureCNN for all classes, with better CoNNText F-1 scores in the case of MRAP and Bradley and slightly lower CoNNText F-1 scores (0.02% - 0.04%) for the other classes (Fig. 5). At 50% and 75% stress levels, CoNNText has a 0.08% - 0.10% lower score for the Abrams class (Fig. 5). The reason is that Bradley has a prominent shape and a distinct silhouette, with a larger turret size and an elevated turret position on the hull. Abrams has a similar shape to Bradley but with a smaller turret and a less prominent shape and silhouette. Distortion affects the Abrams silhouette contours resulting in some Abrams instances being misclassified as Bradley.

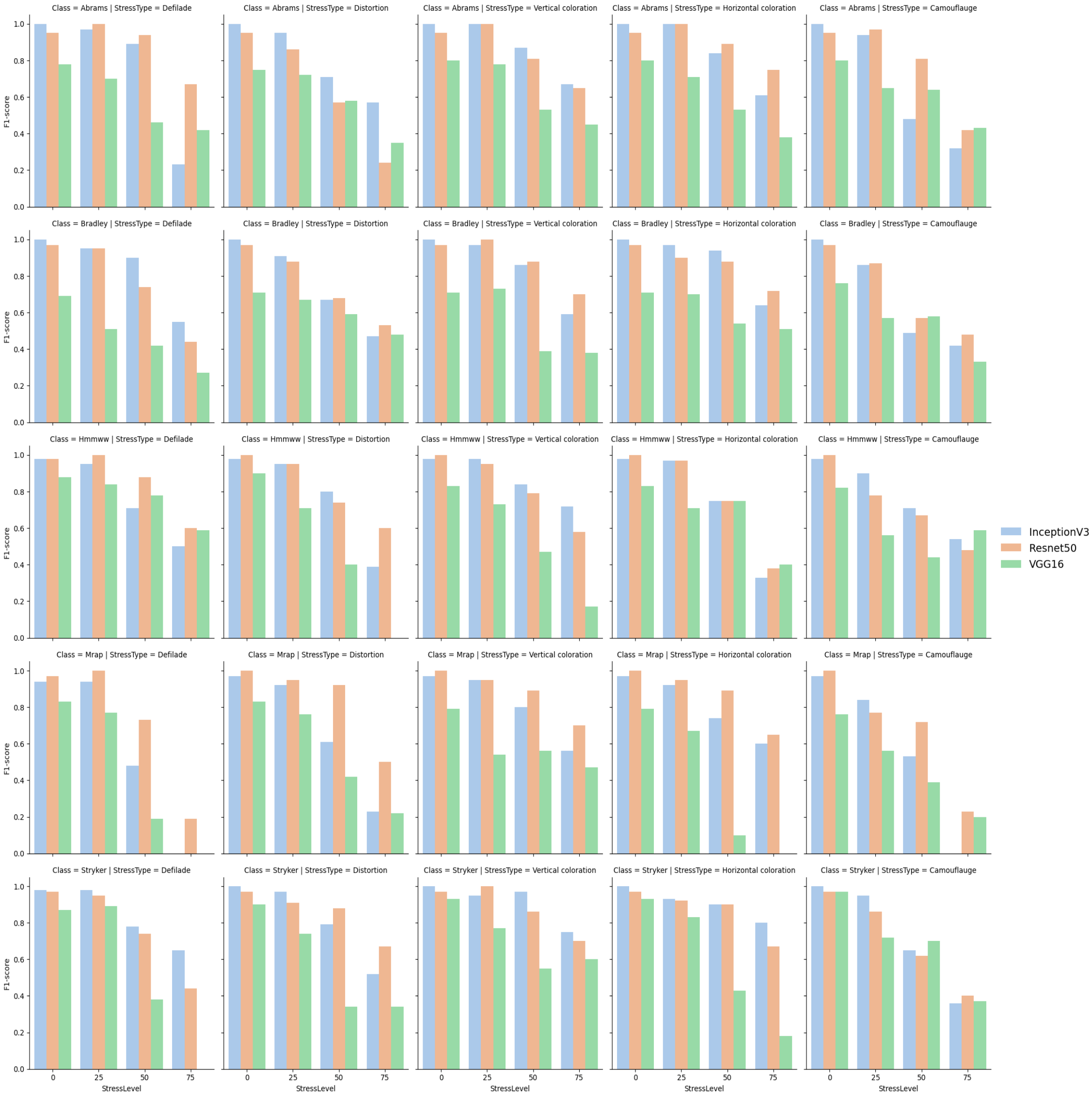

Figure 4: F1-score comparison of the CoNNText models with different base CNNs (Inception-V3, Resnet-50 and VGG-16). The top row to bottom row represents target classes: Abrams, Bradley, HMMWW, MRAP, and Stryker, respectively. The left column to right column represents stressor types: Defilade, Distortion, Vertical Coloration, Horizontal Coloration, and Camouflage, respectively. Within each bar chart: x-axis represents the stress levels - 0%, 25%, 50%, and 75%, respectively; y-axis represents the F1-score. Blue bar: Inception-V3, Orange bar: Resnet50, Green Bar: VGG16.

In the case of the coloration stressor, CoNNText performs better than or the same as the PureCNN (Fig. 5). PureCNN is observed to exhibit a bias towards the Abrams class characterized by low precision and high recall. At the 75% stress level, CoNNText yields better F-1 scores than PureCNN in the case of Abrams and Bradley for horizontal coloration stress since the discrimination between Abrams and Bradley is more dependent on the differences in silhouette shapes. In the presence of high coloration stress which render the PureCNN features less effective, the SC descriptor derived from the silhouette shapes aids in effective target recognition.

For the camouflage stress, the CoNNText outperforms PureCNN in the presence of camouflage stress across all stress levels and classes (Figs. 4, 5, and 6) except for HMMWW (which we discuss further when we analyze the results on a per class basis). The results show the importance of shape-based features (i.e., the SC descriptor) in addressing camouflage stress where RGB image-derived features are rendered less effective.

5.2. Detailed Per Class Analysis

For Abrams, at 25% stress level, Abrams is classified with a high F-1 score by CoNNText compared to PureCNN for all base CNNs and for most stressors except distortion (Fig. 5). A possible explanation is that Abrams, Bradley, and Stryker are the target vehicle classes with mounted guns or turrets on top. The confusion matrix reveals that the incorrect class assigned to Abrams at 25% distortion stress level is mainly Bradley since the difference between these two vehicles is the length of the gun, with the Bradley having a larger gun. Likewise defilade and camouflage are the stressors for which Abrams is misclassfied by CoNNText but only at a higher (75%) stress level.

For Bradley, the CoNNText yields good classification results for Bradley across all stressor types except for defilade at 75% stress level, which is when the turret (Bradley's most distinguishing feature) is obscured (Figs. 4 and 5).

The HMMWW class does not show much improvement with the addition of the shape features via CoNNText since it has no distinguishing shape features (Figs. 5 and 6). Since the HMMWW is a multipurpose vehicle, the HMMWW class exhibits several structural variations making it more difficult to discriminate based on shape features.

The MRAP class shows the most promise for our CoNNText model, due to its unique structure. The MRAP has an unique shape with a high ground clearance and V-shaped hull at the back making the SC descriptor an effective discriminator (Figs. 5 and 6).

The Stryker class shows good results with CoNNText for 50% horizontal coloration stress (Figs. 5 and 6). As in the case of Abrams and Bradley, the turret is the most distinguishing feature of the Stryker class and is captured well by the SC descriptor even in the presence of horizontal coloration stress resulting in higher CoNNText performance.

The proposed CoNNText framework offers an alternative way of improving ATR performance for tasks which require a robust classifier that is immune to environmental noise and distortions. Such a robust classifier would aid a human in real-time decision making. These results demonstrate that low-shot learning with auxiliary information on a domain specific task can aid in target recognition under environmentally stressed conditions. This research also paves the way for further research using additional information like textual data and domain-specific 3D point cloud information to further enhance the robustness of the classifier.

6. Conclusions and Future Work

We proposed a novel ZSL/LSL-based model termed CoNNText for the fusion of CNN-derived feature vectors and SC descriptors as auxiliary information in the context of ATR for military

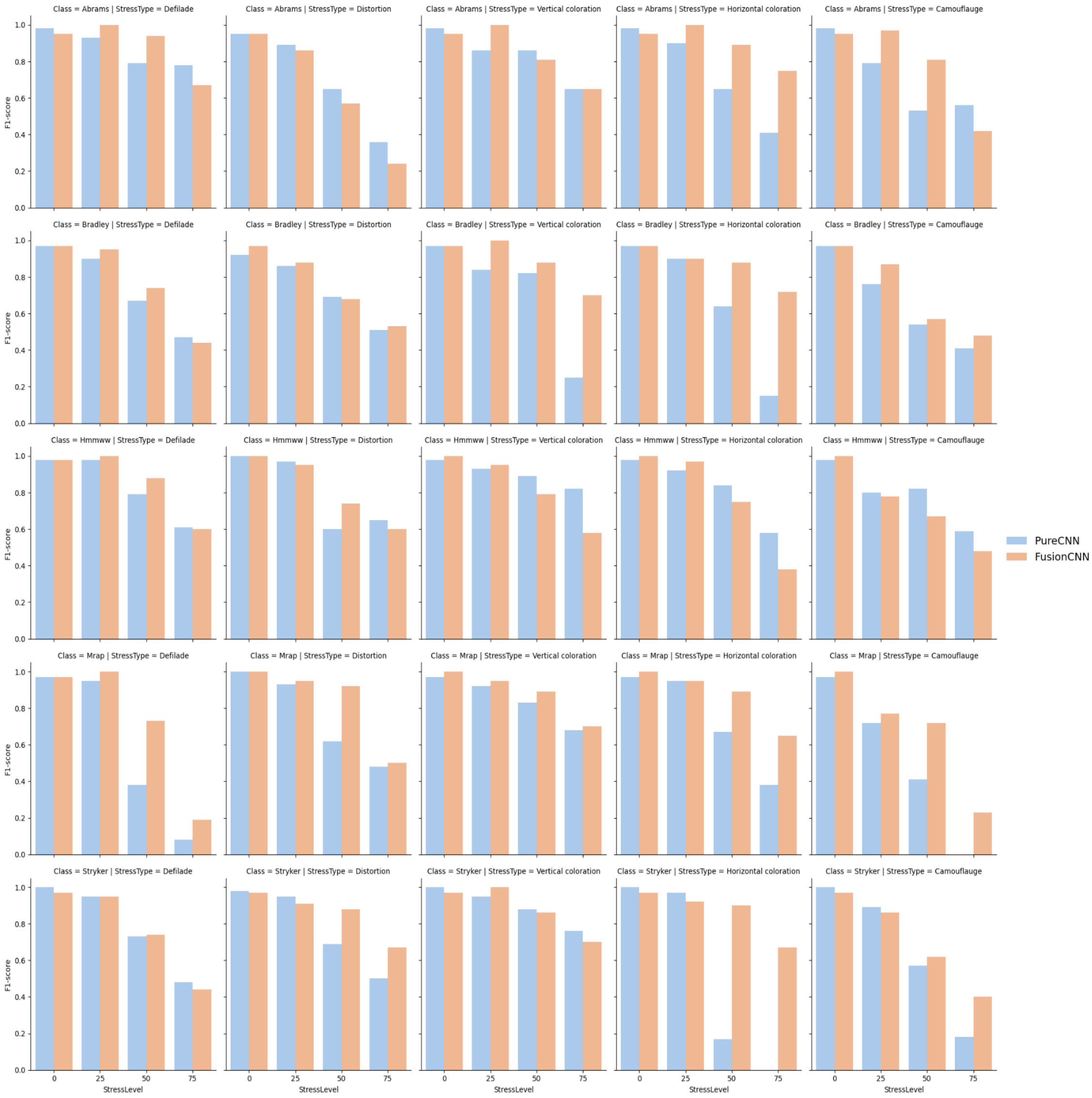

Figure 5: F1-score comparison between CoNNText and PureCNN with Resnet-50 as the base CNN. The top row to bottom row represents target classes: Abrams, Bradley, HMMWW, MRAP, and Stryker, respectively. The left column to the right column represents stressor types: Defilade, Distortion, Vertical Coloration, Horizontal Coloration, and Camouflage, respectively. Within each bar chart: x-axis represents the stress levels - 0%, 25%, 50%, and 75%, respectively; y-axis represents the F1-score. Blue bar: PureCNN, Orange bar: CoNNText.

vehicles under environmental and human-induced stressors.

The CoNNText model showed improved classification accuracy over pure CNN-based approaches and offered an alternative way of improving ATR performance in the presence of environmental and human-induced stressors. The feature fusion was observed to work best in the case of Resnet-50 and least in the case of VGG-16. The CoNNText model yielded improved

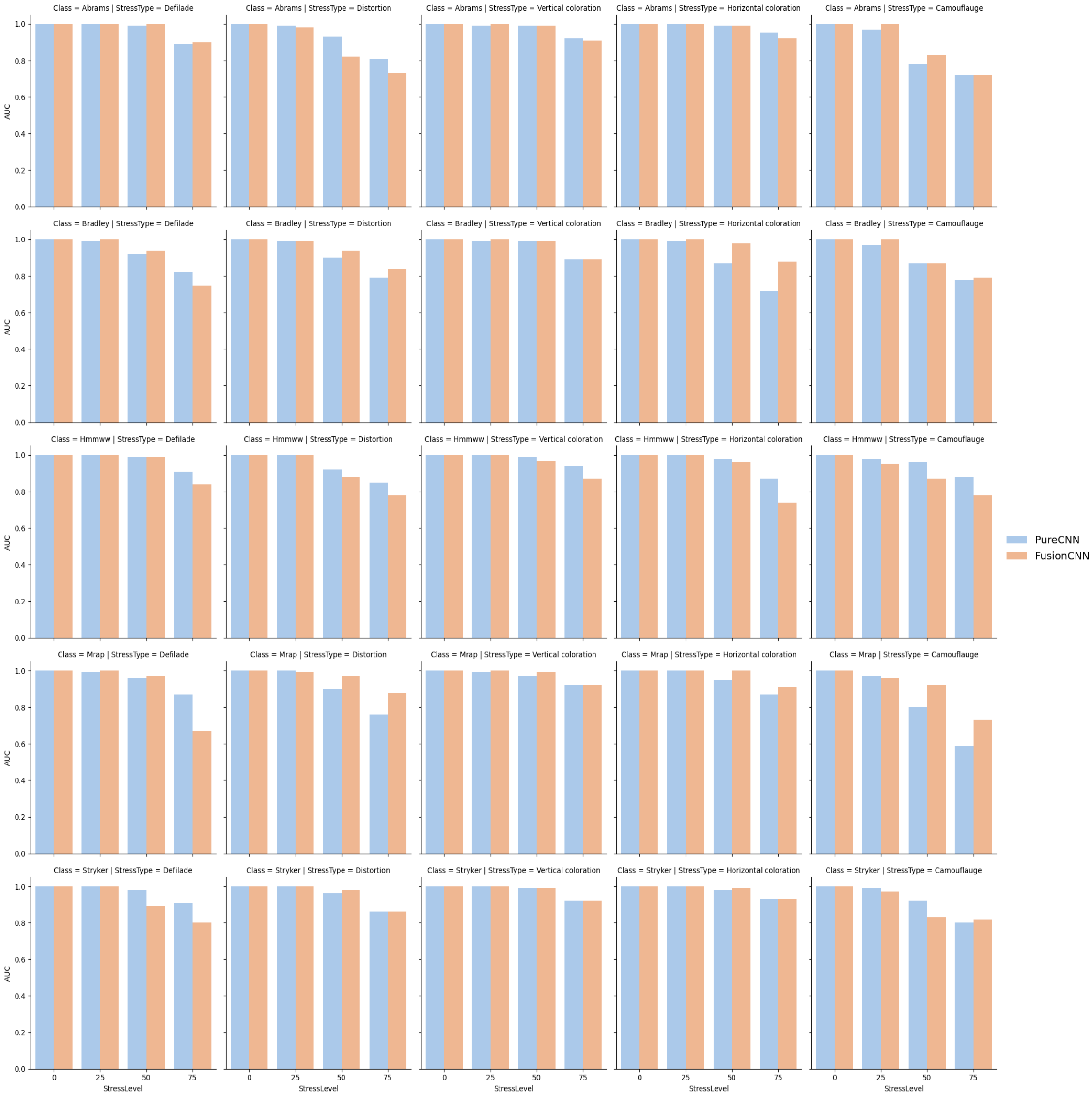

Figure 6: AUC score comparison between CoNNText and PureCNN with Resnet-50 as the base CNN. The top row to bottom row represents target classes: Abrams, Bradley, HMMWW, MRAP, and Stryker, respectively. The left column to the right column represents stressor types: Defilade, Distortion, Vertical Coloration, Horizontal Coloration, and Camouflage, respectively. Within each bar chart: x-axis represents the stress levels - 0%, 25%, 50%, and 75%, respectively; y-axis represents the F1-score. Blue bar: PureCNN, Orange bar: CoNNText

accuracy on a dataset of stressed images from the target classes: Abrams, Bradley, HMMWW, MRAP and Stryker. The improvement in accuracy was 10% - 12% for distortion, camouflage, horizontal coloration and defilade stressors at 50% stress level and 12% for a 50% defilade stressed image. We see a significant increase in accuracy ranging from 36% to 64% on a 75% horizontal coloration stressed image in our CoNNText model with a Resnet-50 base CNN.

Additionally, we experimented with dimensionality reduction techniques such as PCA, ICA, and LDA. We found minor accuracy improvements with ICA and PCA techniques applied to the ResNet-50 architecture. In future, we plan to implement more effective feature selection methods that would yield better classification accuracy. In addition to more effective feature dimensionality reduction, a potential future direction would be to improve the shape descriptor, by focusing on contour extraction rather than instance segmentation. In addition to using shape information, we plan to use attribute and textual information on the target classes, and 3D point cloud estimates as auxiliary information sources in our future LSL/ZSL-based ATR frameworks.

7. Acknowledgement

This research was sponsored by the DEVCOM Analysis Center and was accomplished under Cooperative Agreement Number W911NF-22-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Belongie, S., Malik, J., and Puzicha, J. (2002, April). Shape matching and object recognition using shape contexts, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 24(4), pp. 509-522, doi: 10.1109/34.993558.

[2] Belongie, S., and Malik, J. (2000). Matching with shape contexts, *Proc. Workshop on Content-based Access of Image and Video Libraries*, Hilton Head, SC, USA, pp. 20-26, doi: 10.1109/IVL.2000.853834.

[3] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998, Nov.). Gradient-based learning applied to document recognition, *Proc. IEEE*, Vol. 86(11), pp. 2278-2324, doi: 10.1109/5.726791.

[4] Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017, May). ImageNet classification with deep convolutional neural networks, *Comm. ACM*, Vol. 60(6), pp. 84-90, doi: https://doi.org/10.1145/30653.

[5] Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition, *Proc. Intl. Conf. Learning Representations.*

[6] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

[7] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 38(1), pp. 142-158.

[8] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks, *Proc. Advances in Neural Information Processing Systems*, Vol. 28.

[9] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN, *Proc. IEEE Intl. Conf. Computer Vision*, pp. 2961-2969.

[10] Lampert, C.H., Nickisch, H., and Harmeling, S. (2014). Attribute-based classification for zero-shot learning of object categories, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 36(3), pp. 453-465.

[11] Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2018). Zero-shot learning — A comprehensive evaluation of the good, the bad and the ugly, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 41(9), pp. 2251-2265.

[12] Socher, R., Ganjoo, M., Manning, C.D., and Ng, A. (2013). Zero-shot learning through cross-modal transfer, *Proc. Advances in Neural Information Processing Systems*, Vol. 26.

[13] Vinyals, O., Blundell, C., Lillicrap, T., and Wierstra, D. (2016). Matching networks for one-shot learning. *Proc. Advances in Neural Information Processing Systems*, Vol. 29.

[14] Finn, C., Abbeel, P., and Levine, S. (2017, July). Model-agnostic meta-learning for fast adaptation of deep networks, *Proc. Intl. Conf. Machine Learning*, pp. 1126-1135.

[15] Debroux, P.S. (2022), *Analysis Methodology of Image Classifiers in Stressed Environments*, US Army DEVCOM Analysis Center Technical Report.

[16] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 248-255.

[17] J.R. Wilson (2019). *Artificial intelligence (AI) in unmanned vehicles*, https://www.militaryaerospace.com/.

[18] Xu, K., Ba, J., Kiros, R., Cho, K. Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015) Show, attend and tell: Neural image caption generation with visual attention, *Proc. Intl. Conf. Machine Learning*, Vol. 37. pp. 2048–2057.

[19] Chao, W-L et al. (2016) An empirical study and analysis of generalized zero-shot learning for object recognition in the wild, *Proc. European Conference on Computer Vision*, pp. 52-68.

[20] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception architecture for computer vision, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2818-2826.

[21] Bourlard, H., and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition, *Biological Cybernetics*, Vol. 59(4-5), pp. 291-294.

[22] Perkins, D.N., and Salomon, G. (1992). Transfer of learning. *International Encyclopedia of Education*, Vol. 2, pp. 6452-6457.

[23] Pan, S. J., and Yang, Q. (2009). A survey on transfer learning, *IEEE Trans. Knowledge and Data Engineering*, 22(10), pp. 1345-1359.

[24] Goodfellow, I.J., Bengio, Y., and Courville, A. (2016) *Deep Learning*, MIT Press.

[25] Gao, J., Li, P., Chen, Z., and Zhang, J. (2020); A survey on deep learning for multimodal data fusion, *Neural Computing*, Vol. 32(5), pp. 829–864, doi: https://doi.org/10.1162/neco_a_01273.

[26] Ramachandram, D., and Taylor, G.W. (2017) Deep Multimodal learning: a survey on recent advances and trends, *IEEE Signal Process. Mag.*, Vol. 34, pp. 96–108.

[27] Everything you need to know about Few-Shot Learning https://blog.paperspace.com/few-shot-learning/