Multimedia Tools and Applications (2023) 82:39435–39457
https://doi.org/10.1007/s11042-023-15056-y

Connecting national flags – a deep learning approach

Theofanis Kalampokas[1] · Dimitrios Mentizis[1] · Eleni Vrochidou[1] · George A. Papakostas[1]

Received: 9 December 2022 / Revised: 5 February 2023 / Accepted: 28 February 2023 /
Published online: 29 March 2023

Abstract

National flags are the most recognizable symbols of the identity of a country. Similarities between flags may be observed due to cultural, historical, or ethical connections between nations, because they may be originated from the same group of people, or due to unrelated sharing of common symbols and colors. Although the fact that similar flags exist is indisputable, this has never been quantified. Quantifying flags' similarities could provide a useful body of knowledge for vexillologists and historians. To this end, this work aims to develop a supporting tool for the scientific study of nations' history and symbolisms, through the quantification of the varying degrees of similarity between their flags, by considering three initially stated hypotheses and by using a novel feature inclusion (FI) measure. The proposed FI measure aims to objectively quantify the overall similarity between flags based on optical multi-scaled features extracted from flag images. State-of-the-art deep learning models built for other applications tested their capability for the first time for the problem under study by using transfer learning, towards calculating the FI measure. More specifically, FI was quantified by six deep learning models: Yolo (V4 and V5), SSD, RetinaNet, Fast R-CNN, FCOS and CornerNet. Flags' images dataset included flags of 195 nations officially recognized by the United Nations. Experimental results reported maximum feature inclusion between flags of up to 99%. The extracted degrees of similarity were subsequently justified with the help of the Vexillology scientific domain, to support research findings and to raise questions for further investigation. Experimental results reveal that the proposed approach and FI measure are reliable and able to serve as a supporting tool to social sciences for knowledge extraction and quantification.

Keywords Deep learning · Computer vision · Cultural heritage · Flags similarity · Similarity measure · Artificial intelligence · National flags · Vexillology

✉ George A. Papakostas
gpapak@cs.ihu.gr

[1] MLV Research Group, Department of Computer Science, International Hellenic University, 65404 Kavala, Greece

1 Introduction

National flags are unique symbols that represent an ideal or an idea, aiming through simple designs and combinations of elements, shapes, and colors, to capture the history, culture, and values of countries. Meaningful symbolisms and combinations of colors are used, so as for each flag to be either distinctive or intentionally related to another. Therefore, similarities are inevitable, due to national overlapping, meaning that several counties share the same history, culture, or religious ideals, or due to design principles overlapping, meaning that designers share the same definitions of shapes and colors, and follow the same simple design rules that imply finite lines and basic colors.

The similarity between flags has been investigated in the literature [1, 10, 17, 25, 32]. However, the focus was mainly on grouping similar flags and finding links between nations and the structure of their symbols. Knowledge discovery from flag images and web documents has been investigated in the past [32]. The authors extracted features from both images and text, such as the frequency of occurrence of a word, color, shape, and texture. Then, they defined similarities between nations by calculating the Euclidean distance between feature vectors, by using the Self-Organizing Map (SOM) algorithm. In [17], machine learning algorithms were used for national flag classification. Features were extracted from flag images and were used to comparatively evaluate the classification performance of a Multilayer Perceptron, a Classification And Regression Tree (CART), and a Decision Tree Classifier (C4.5) to predict the religion and landmass of the country. National flag classification has also been conducted by Akhand et al. [1]. The authors extracted features from flag images and correlated them with religion, government, and region of countries, by using the C4.5 algorithm. It should be noted that the exploitation of deep learning algorithms for feature extraction from flag images has not yet been reported in the literature. Moreover, even though the similarity between national flags has been confirmed and interpreted, it has never been quantified.

Machine learning [3, 24] and deep learning [5, 6] methods have been extended from classical applications and have also been applied to applications related to cultural heritage. Machine learning algorithms have been proposed to make statistical analyses of different kinds of cultural heritage data based on classical methods like regression, classification, or clustering [1, 17]. These approaches aim to extract information from symbols regarding the lives and events of descenders in each nation and, thus, fill knowledge gaps in their national history. One of the recent contributions of machine learning to cultural heritage was presented by López-García et al. [22], where a Random Forest (RF) classifier was used to classify ceramic artifacts based on their chemical composition. Polak et al. [26] used hyperspectral images of artworks to train an SVM model to identify and classify pigments with the aim to preserve authentication of cultural findings in art. The main difficulty in implementing machine learning methods in cultural heritage related applications is the reported lack of available datasets, which are either small or sparse in amount and structure or not public, leading to less concrete conclusions.

Recently, deep learning methods and more specifically Deep Neural Networks (DNNs) have also been applied to cultural heritage related applications. DNN models can assimilate a large amount of data, even if it is sparse, and can extract valuable information with the appropriate implementation. Small datasets can therefore be handled by DNN models by either using transfer learning and/or data augmentation techniques. Moreover, DNNs can gradually learn through each layer increasingly complex data representations. Therefore, DNNs have been used extensively in applications of computer vision and natural language processing, and subsequently adapted to cultural heritage since artworks and documents are the main objects preserved from the past that can be used to extract knowledge for dark periods in history. Most deep learning applications in cultural heritage are associated with documents and ancient characters recognition, since there are many preserved historical texts, either in print or in murals, petroglyphs, etc., in several countries globally [11, 23, 33]. Applications of deep learning in cultural heritage can also be found in artworks [13, 28, 30].

To this end, this work exploits the benefits of deep learning methods to quantify for the first time the similarity between national flags. National flag images are used as input data to extract and locate overlapping features between them by using six deep learning models. Experimental results are matched with Vexillology facts to interpret and verify the appearance of image features of one flag inside another. The aim of this work is to make underlying symbolisms of flags more accessible and significant by providing deeper insights into the history behind them. The proposed approach can serve as a complementary tool to social sciences for analytical and didactical purposes. The main contributions of this work can be summarized as follows:

- For the first time, the correlation between national flags is quantified.
- For the first time, deep learning is employed to correlate nations based on their flags. The challenge is to investigate whether popular deep learning models built for other applications are capable of being successfully adapted for the first time to the problem under study by using transfer learning.
- A novel optical feature inclusion measure is proposed and applied to quantify the similarities between national flags.
- For the first time, the Vexillology discipline is considered in a computer vision application to support the experimental findings, providing historical, political, and social insights that correlate nations.
- For the first time, a benchmark flags dataset is provided to the research community comprising of 195 classes. The number of classes is higher than already well-known benchmark datasets such as COCO.

The rest of this paper is organized as follows: In Section 2 the history of national flags is briefly stated focusing on well-known correlations between them. Section 3 presents related work of deep learning in cultural heritage applications. Materials and methods are presented in Section 4, while in Section 5 the experimental results are reported. Section 6 concludes the paper and proposes future work directions.

2 History of national flags

Flags' origins date back to ancient times, first appearing in the East. There were mainly used in warfare, to identify friends or enemies. Flags had multiple symbolisms: could indicate captivity, punishment, defeat, or used just for signaling. The flag was the first object of attack in a battle, and its fall would indicate defeat. In Europe, national flags were first introduced in the Middle Ages and Renaissance as accepted symbols of countries.

The basic design attribute of a flag is its color. Originally, color was bonded with a family dynasty or an empire. For national flags, color may refer to nature or may have national symbolic meanings that vary based on the ideals of each country [10, 20]. Therefore, blue may represent the sea, the water, or the blue sky, e.g., the Somalian and Greek flags, or can signify determination, liberation, tranquillity, and calmness, e.g., the flag of Kazakhstan. Red color refers to revolution, hardiness, and bloodshed, e.g., the flags of Albania, China, and Turkey. Green refers to nature and agriculture, thus, it is met in many agricultural oriented countries such as Brazil and Zambia. The flag of the United States shares the same colors with the flag of United Kingdom, due to its diplomatic relations with Britain, shared history, language, and religion. Similarly, flags' designs stem from the history, culture, or religion of each country. Therefore, common design elements or patterns appear on flags of countries that share the same history and culture, including basic shapes such as stars, stripes, and crosses. More specifically, flags can be classified based on their design into two main categories: cross flags and stripe flags, along with their variations, as presented in Fig. 1. Cross flags have the cross symbol placed in the center, while in the Scandinavian flag, the cross is placed closer to the hoist than to the fly; saltire flags have the cross couped; stripe flags can be subcategorized based on the direction and the number of stripes: bicolor flags have two horizontal or vertical stripes of two different colors; tricolor flags have three horizontal or vertical stripes of three different colors; triband flags have three horizontal stripes of two different colors. In general, similarities in flags may be coincidental; however, most times based on Vexillology, they represent shared connections between countries.

Common origins between flags can be tracked in general flag families that share the same design attributes. Flag families are linked by geographical position or by common history, culture, and traditions. For example, countries displaying the Christian cross on their flag were influenced by the Crusades, when this cross symbol was first introduced. Some of these countries are England, Scotland, Ireland, Denmark, Greece, and Switzerland. Of special attention is the flag of United Kingdom, namely the Union Jack, which incorporates the three crosses of the flag families of St. George of England, St. Andrew of Scotland, and St. Patrick of Ireland, as illustrated in Fig. 2. All three component flags share similar heraldic cross designs, influenced by their shared history.

Long wars for independence are behind similarities of some striped flags, forming another family of flags associated with liberty and republican governance. The horizontally striped flag of Netherlands, which was used during the war for independence from Spain, inspired France, who adopted the same colors on vertical stripes after the French Revolution. The same inspiration is behind the Dutch flag, followed by the Russian tricolor flag which, in turn, formed the basis of many national flags of eastern European countries. Figure 3 illustrates similar influenced national flags.

Based on the above, it is obvious that national flags encode rich information about each country. In fact, there is a separate field of science, namely Vexillology, that studies symbols

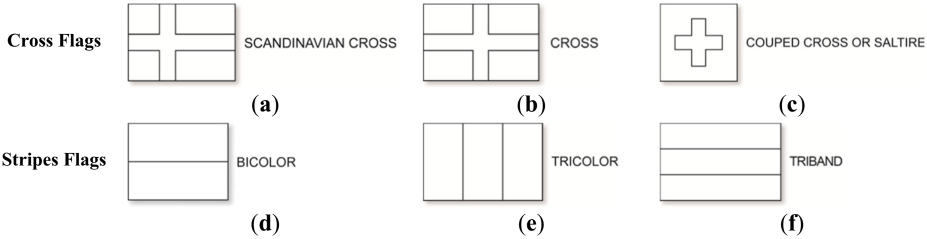

Fig. 1 Cross and Stripes flag categories: **a** Scandinavian cross; **b** Cross; **c** Saltire; **d** Bicolor; **e** Tricolor; **f** Triband

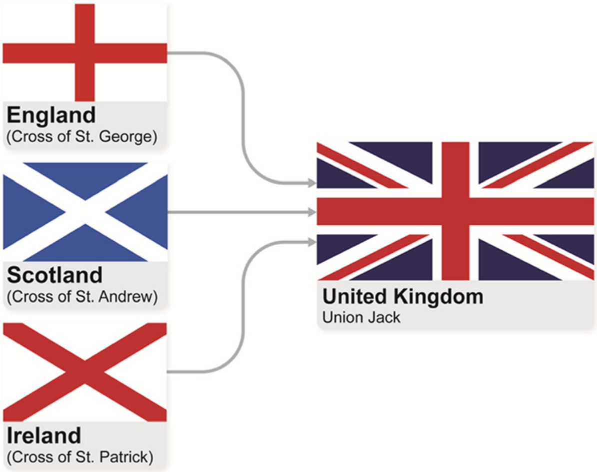

Fig. 2 Family flag of United Kingdom (Union Jack)

and history behind the formation of flags. In this work, Vexillology assumptions regarding the connection of countries based on their national flags are supported by computer vision. More precisely, deep learning methods are used to quantify and generalize the various data embodied in national flags, proving that machines can provide unbiased decisions.

3 Related work

Deep learning has been established as a powerful tool in many scientific sectors due to its ability to extract high-level features for complex pattern recognition problems. Cultural heritage has also exploited the benefits of deep learning, especially for image and Natural Language Processing (NLP) related applications, such as for automatic image captioning that combines both computer vision and NLP [2]. Preservation and diagnostics of cultural heritage findings, e.g., paintings, sculptures, documents, and artworks, are crucial to determine the historical status of findings and extract the missing knowledge. In this scientific field, Belhi et al. [7] conducted related research. His research team collected 10,000 artworks in the form of 2D images and trained a set of state-of-the-art convolutional neural networks (CNN), Visual Geometry Group (VGG) 16, VGG19, and Residual Neural Network 50 (ResNet50) with transfer learning, to make conclusions regarding their creation year, creator and genre, reporting a prediction accuracy of 85%. Sabatelli et al. [28] used four DNNs with two different

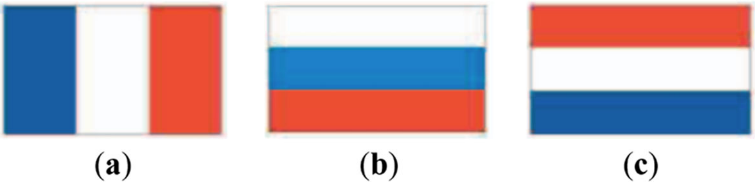

Fig. 3 Similar influenced national flags: **a** French flag; **b** Dutch flag; **c** Russian flag

approaches known as off-the-shelf classification to recognize materials, authors, and artistic categories of artworks. Jboor et al. [15] implemented a framework based on three artwork image datasets. Global features were extracted from each image of the datasets based on VGG16, VGG19, and ResNet50. Researchers first applied Scale-Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF) algorithms and used K-means clustering to categorize the extracted features. Second, they trained a Generative Adversarial Network (GAN) model to inpaint the damaged parts of artwork images of the same category, based on the clustering results.

An alternative deep learning application of cultural heritage in the archaeological domain was presented by Gallwey et al. [12]. The authors proposed a pre-trained CNN called DeepMoon with the ability to process Digital Elevation Models (DEMs) images of terrain surfaces to detect holes indicating subjected archaeological findings. In the same direction, Sharafi et al. [29] developed an application that exploited surface images from airplanes combining deep learning models and other machine learning methodologies to predict possible locations of archeological buried findings. The performance of methodologies that were applied in that problem achieved AUC (Area under Curve) of up to 98% in correct prediction of locations. Assael et al. [4] used a deep learning neural network, DeepMind, to recognize text from an ancient Greek epigraphy, to extract and restore the content from images. A didactic application was proposed by Bongini et al. [9]. A neural network was trained to reply to questions about image content such as artwork, providing the user with useful information in terms of general knowledge. The latter could potentially be exploited as a museum guide for the historical analysis of art creations.

All above mentioned implementations form only a small indicative part of the numerous applications of deep learning in natural heritage. Based on the targets of the referenced applications, it can be concluded that most efforts focus on the preservation of cultural heritage and the extraction of knowledge. However, the extraction of knowledge is only limited to typical identification information (age, area, materials, etc.) and not to the deeper analysis of cultural information which may lead to a demystification of historical, social, political, religious events and behaviors that affected certain nations and subsequently influenced others. Moreover, there is no reported research in the bibliography, as far as our knowledge, using deep learning in natural heritage towards correlating information and quantifying the relationship between different civilizations.

In this direction, this work proposes an approach to fill the identified research gap, by using deep learning to quantify the connections between nations based on data encoded in their flags. In this work, flags are deeply investigated to extract features and quantify inclusions, exploiting flag elements and colors that are known to reflect the cultural heritage and history of nations through historical symbolisms related to physical/geographical characteristics of nations, battles, liberation, natural resources, and more.

4 Materials and methods

4.1 Proposed methodology

Computer vision methods are adopted for the interpretation of feature inclusions between national flags. Figure 4 illustrates the flow of the proposed methodology. Initially, a CNN-based object detection model is trained with a national flags image dataset. Then the model is

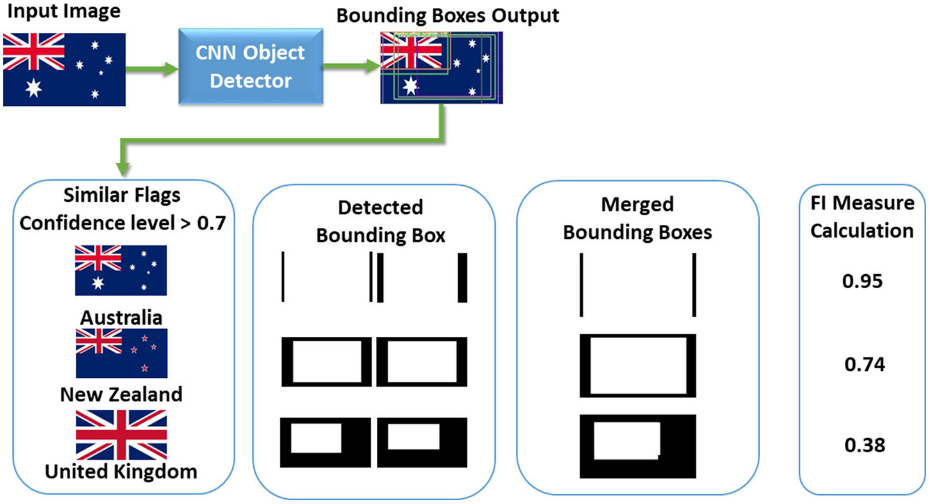

Fig. 4 Block diagram of the processing steps of the proposed methodology

applied to localize and classify bounding boxes of national flags into other flag images. The input to the CNN model is a flag image; at this step, the performance of six different deep learning models is investigated. The output of the model is the bounding boxes with the model's confidence. After merging the extracted bounding boxes (having confidence level > 0.7) that include similar patterns, the proposed feature inclusion measure is calculated as the ratio of the pixel area of the predicted bounding boxes with similarities in an image flag with another, divided by the whole flag pixel area. The illustrated example of Fig. 4 refers to the investigation of the similarity of the national flag of Australia with other flags e.g., New Zealand and United Kingdom (UK). For the latter case, the scaled pattern of the flag of UK is detected in Australia's flag on the upper left side. This detection is used to calculate the inclusion measure of the flag of UK in Australia's flag. Six different CNN object detectors were studied towards finding the most appropriate for dealing with this challenging task.

Numerical results that indicate high correlations between the flags of nations derived from this process are subsequently interpreted and verified through the Vexillology domain. Therefore, historical connections between flags were verified by artificial intelligence methods and for the first time these connections were numerically quantified by the proposed inclusion measure.

Therefore, an analysis of national flag images is conducted to identify, measure, and evaluate the correlation between flags, by combining the research fields of computer vision and Vexillology. It should be noted here that no assumptions were made during the simulation phase of this work. Hypotheses, however, were made and verified from the testing scenarios:

- **Hypothesis #1.** There are flags of different nations that are similar to each other.
- **Hypothesis #2.** The proposed inclusion measure is able to measure the overall similarity between two flags.
- **Hypothesis #3.** The proposed models are able to detect multiscale patterns inside the flags.

In what follows, object detection models, the feature inclusion measure used in this work, and data preparation steps, are presented.

4.2 Proposed object detection models

Object detection is a field of computer vision that has experienced a huge rise in the last years with numerous applications that aim to detect objects and areas of predefined categories inside images. This fact gave a significant impact on the Deep Learning domain leading to the evolution of CNNs' architectures. Classical object detection models can be divided into two main categories: *one-stage* and *two-stage*. Both categories compute a large number of anchors on the image and predict their category. After the classification phase, the coordinates of the positive (correct classified) anchors are refined and proposed as the detection results. The process is repeated several times to optimize the localization and classification of objects in the image. On the one hand, two-stage detectors repeat more than one-stage detectors and the whole image process takes place in two parts (regression – classification), which justifies their high computational demands. The above scheme is followed in Region-Based CNN (R-CNN), which is a basic two-stage object detection model. On the other hand, one-stage object detection models have become competitive over two-stage detectors with the introduction of You-Only-Live-Once (YOLO) and Single-Shot Detector (SSD), being able to apply classification and localization in one regression process. One-stage detectors produce a large number of anchor boxes (like two-stage) with different scales over image feature maps that are directly classified and refine their coordinates through multi-scale convolutional layers in a single structure. The latter mechanism endows in one-stage detection models the advantage of high computational efficiency.

Despite the great impact of both categories, anchor-based object detection methodologies also have disadvantages. Anchor-based detectors have a strong link between the detection performance and the number, aspect ratios, and size of anchor boxes, which make them unstable in problems where objects are of various shapes and scales. Another disadvantage of anchor-based object detectors is that a large number of anchor boxes are produced in order to generalize (more than 100 k in RetinaNet) while in the end only a small number of them are kept; thus, redundancy occurs. A solution is provided by anchor-free mechanisms, adopted by object detection CNNs. Anchor-free object detectors have differentiated from anchor-based models with two different approaches: *center-based* and *keypoint-based*. Center-based detectors compute object boxes based on the center of positive regions and then predict four distances from the object box's boundaries. These kinds of anchor-free detectors are similar to anchor-based detectors with the difference that the model does not predict boxes' coordinates but four values that represent the distance from the box center. Center-based detectors have achieved similar performances to anchor-based detectors, however with less computational demands. Keypoint-based detectors first locate several pre-defined or self-learned keypoints and then produce bounding boxes based on keypoints' distances to detect bounding boxes around objects. Anchor-free object detectors have made a significant contribution to the field of computer vision.

The challenge is to investigate whether popular deep learning models that have been built for other applications are capable of being successfully applied for the first time to the problem under study by using transfer learning. In the current study, six object detection models from all four categories presented above are proposed and used to extract feature inclusions of one national flag in another in the form of a bounding box. The selected models are: YoloV4 [8], YoloV5 [16], SSD [21], RetinaNet [19], Fast R-CNN [14], Fully Convolutional One-Stage Object Detection (FCOS) [31] and CornerNet [18].

4.2.1 Yolo

Yolo was introduced by Redmon et al. [27] as the first object detection model that connected the procedure of regression boxes and classification of them in an end-to-end CNN. Yolo series models are some of the dominant models of one-stage object detectors and some of the most used models in computer vision applications. YoloV1 was the first anchor-free object detector, which exploited the center of the bounding box to refine the box's coordinates. To handle the disadvantage of YoloV1 to detect small objects, versions YoloV2 and V3 used an anchor-based mechanism with the addition of 53 layers in the Darknet backbone to achieve three different sizes of feature maps. The architecture of Yolo consists of three phases: the backbone which is a CNN that extracts multi-scale feature maps from an input image, the neck which is a series of layers that apply feature fusion from low- and high-level feature maps that are produced to extract more contextual information, and the head which predicts the coordinates and the class of a given object. YoloV4 [8] proposed a new backbone architecture CSPDarknet53 (Cross-Spatial -Partial DarkNet53), with Spatial Pyramid Pooling (SPP) as an additional module, PANet path-aggregation as neck, and finally, the head which remained the same as in YoloV3. Considering the above-mentioned additions, YoloV4 achieved a 10% and 12% increase in Average Precision (AP) and Frames per Second (FPS), respectively. In YoloV5, Feature Pyramid Network (FPN) and Path Integral Based Convolution (PAN) network structures are used in the neck to achieve better performance in the same processing time for training and inference stages. The rest of the model remained the same with YoloV4. To date, it is considered the state-of-the-art model in object detection problems, holding the state-of-the-art performance of 1666 FPS in 640 × 640 pixels image size, with a batch size of 32 in Tesla v100 Graphics Processing Unit (GPU).

4.2.2 SSD

SSD model was introduced by Liu et al. [21] and since then it has proven its efficiency in object detection problems regarding speed and accuracy reporting 76.9% mAP (mean Average Precision) in the COCO dataset. Its architecture can be divided into two parts: the backbone where VGG16 is used without the final fully connected classification layer to extract multi-scale feature maps from an input image, and the SSD head which is another set of fully convolutional layers with different sizes with a final output of box coordinates and the class of the box. With the multi-scale structure from the backbone to the head, SSD was proven robust in handling various scales of objects.

4.2.3 RetinaNet

RetinaNet was introduced by Lin et al. [19] as the first one-stage object detection model that outperformed all two-stage object detectors of that time. The architecture of RetinaNet is divided into two parts: the backbone, which is a ResNet for feature extraction, and an FPN network on top, where each pyramid level consists of two subnetworks one for classification of the object classes and one for regression of object bounding boxes. RetinaNet was proposed with a new loss function that could handle positive/negative sampling in a different way than conventional object detection models. RetinaNet loss function decreases to zero with the higher classification confidence, which leads to down-weight easy examples, leaving the space in the network to fit in more complex samples. The above mechanism combined with the huge amount of anchor boxes produced during training gave RetinaNet the capability to emerge among the state-of-the-art models. The main contribution of RetinaNet architecture and mechanism was the ability to outcome the limitations of one-stage detectors, and the diversity of foreground-background classes in data, in combination with dense anchor sampling.

4.2.4 Fast R-CNN

Fast R-CNN is also a core model in two-stage object detectors and was introduced by Girshick [14]. It consists of a backbone with the addition of Region Proposal Networks (RPNs) and region-wise prediction network R-CNN. Taking an image as input, a feature map is fabricated and fed into the Region of Interest (RoI) Pooling layers in order to extract RoIs in the form of feature vectors. These vectors are fed into a structure of fully connected layers with two outputs: the class through softmax activation of its region and a four values vector with the bounding box coordinates. Fast R-CNN attracted a lot of attention from the research community and has been used in various applications.

4.2.5 FCOS

FCOS is an anchor-free one-stage object detection model introduced by Tian et al. [31]. FCOS uses FPNs where in each pyramid level a classifier and regression modules are attached with an extra branch that calculates center-ness. The center-ness value is then multiplied by the classification score of the predicted bounding box to eliminate false positives in Non-Maximum Suppression (NMS) procedure. From the computed feature maps of the FPN network, each point is correlated to the ground truth area and if it belongs to the ground truth box then it is kept as a positive sample; if it belongs in more than one box then it is treated as ambiguous. From the positive points, four distances from each box side are passed through the regression in contrast with anchor-based object detectors, which process four axis points. FCOS object detection model outperformed in at its time all one-stage anchor-based and two-stage object detectors for the COCO dataset.

4.2.6 CornerNet

CornerNet [18] is another anchor-free object detector from the category of key-point-based detectors and belongs to two-stage object detectors. It predicts objects through the process of keypoints in the form of heat maps where its keypoint corresponds to the left-top and right-bottom corner of the object bounding box. More analytically, CornerNet adopts the Hourglass network, which is used to estimate human poses with keypoints, such as palm, head, and hands. The Hourglass network is then connected with two prediction modules with its own corner pooling module; one is for the left-top corner and the other for the right-down. Then corner grouping is accomplished by the prediction of an embedding vector for each corner such that the distance calculation between the embedding vectors with small values will belong to the same object and, thus, the predicted boxes are formed. Finally, for predicting more accurate boxes, an offset is calculated to adjust the corner location. CornerNet managed at its time to outperform most of the two-stage and one-stage object detectors for the COCO dataset.

4.3 Proposed feature inclusion measure

Two primary types of similarities have been observed between national flags. The first one is the instances where one national flag is contained in another. For example, the Union Jack appears unchanged on the upper left corner of the national flag of Australia. The other type of similarity is the instances where two flags are similar in their whole area. For example, the national flags of Jordan and Palestine, have the same design and colors with small differences in their contained elements; they are similar except Jordan has a white seven-pointed star in the middle of the red triangle.

With the above facts, a measure is proposed to quantify the feature inclusion between flags, where the pixel area of the predicted bounding box in an image flag is divided by the whole flag pixel area. The proposed feature inclusion measure is calculated as follows:

$$FI = \frac{\text{Pixel area of predicted flag}}{\text{Pixel area of the flag}} \tag{1}$$

The calculated FI value is then normalized to belong to the interval [0, 1] and corresponds to the ×100% of features of one flag included in another.

4.4 Dataset

In this work, experiments are designed by following a well-defined and finite number of steps. These steps include: (1) dataset collection, (2) data preparation, (3) model selection, (4) model training, and (5) performance evaluation. To provide a fair comparison between the performance of the selected models, each model is trained with the same dataset with minor differences in models' requirements only when this is necessary, e.g., the input image size. For the dataset collection, the national flags of all modern nations of the world were collected from the internet. Because of political and historical issues, there are often conflicts over which states are considered nations. For this reason, the images of the national flags of the 195 nation-states recognized by the United Nations have been considered.

Data preparation includes data augmentation, and application of the following techniques:

- Gaussian noise.
- Brightness manipulation.
- Exposure manipulation.
- Image re-scale.

Two steps of Gaussian noise were applied to the original images. Moreover, the original images were resized to half of their length. Since each national flag has a different aspect ratio, the original dataset had a fixed length of 640 pixels for each flag's length and there was a variation in heights. To deal with these design aspects and to prevent changes in the aspect ratio of each flag, simultaneously, the size of the images was calibrated to half of the original length, by sequential subtraction of 32 pixels. The height was recalculated by multiplying the new length with the quotient of the aspect ratio of the original image. The main goal of the dataset augmentation was to stress the models to detect and extract the same features on different scales. Based on the applied augmentation methods, the final flags' image dataset included 22,180 images for training and 1989 images for evaluation. With the above generated dataset of national flags, six object detectors (seven model architectures) were trained. After the training phase, 195 national flag images were applied to each model to extract predicted bounding boxes that belong to other flags.

In what follows, the experimental results are presented from two different perspectives. First, the analysis and the conclusions from the experiment results are presented to support Vexillology assumptions. Next, the experimental results are presented and analyzed from a technical point of view, where the performance of each model is evaluated.

5 Results and discussion

5.1 Vexillology correlations

All models were trained with the same final flags' image dataset of 22,180 images, belonging to 195 classes. The similarity measure between flags was calculated in terms of the proposed FI measure. Experimental results have been categorized based on the 14 defined flag types according to the Vexillology domain defined in Table 1. From the 195 national flags, similar flag pairs have emerged. In what follows, 26 flag pairs with the maximum inclusion measures are presented, categorized by their type according to Table 1. More specifically, maximum similarities were found between flags belonging to six of the representative flag designs that are marked in bold in Table 1.

Tricolor is the dominant pattern among national flags, as seen in Table 1. Triband designs are symbolizing republicanism, liberty, or revolution and are also common patterns. Therefore, stripped flags including the most popular types of Tricolor, Tribar, and Bicolor were found as the most correlated and dominant patterns cumulatively. Cross and Scandinavian cross are also a broader category and high inclusions were found between them, since historically crosses have influenced many nations, such as the Union Jack. Finally, Canton flags were also found similar, as being present in multiple flag designs due to the fact that a canton usually means unity of the nation, colorized with blue, white, or red.

Table 1 Flag type's categorization

Flag Type	Number of flags
Tricolor	**54**
Plain with emblem	20
Triangle	18
Tribar	**16**
Bicolor	**16**
Canton	**14**
Bend	8
Scandinavian cross	**5**
Cross	**4**
Quartered	4
Serration	2
Bordered	2
Saltire	2
Other types	30

Popular patterns with higher correlations are marked in bold

5.2 Testing scenarios

In this work, 10-fold cross-validation was used to estimate the models' performance. All information regarding the models' parameters, is included in Table 2. The models' global parameters were determined after trial and error; for the selected parameters the training performance was robust for all models towards a fair comparison between them. Table 3 includes the models' optimizers' parameters and losses. Furthermore, the specifications of the desktop computer where all experiments were conducted are summarized in Table 4.

In the following Tables, the flag pairs are presented based on the object detection results and the feature inclusion measure (Flag1 – Flag 2: FI of Flag 1 included in Flag 2). It should be noted that FI reached up to 99%, numerically justifying the correlations resulting from Vexillology.

Table 5 includes the Tricolor flags feature inclusion results. According to Table 5, based on feature inclusion results and Vexillology, Chad and Romania share the same design and colors in their national flags. The blue color in both flags represents the sky and the red color signifies independence and bloodshed, whereas the third color has a different meaning for both countries. Romania's flag third color is orange, which stands for hard work, while Chad's flag third color is gold, which stands for the desert and the sun.

Senegal and Mali share common colors and designs where the red color stands for the blood and sacrifices for independence, yellow stands for wealth in both countries, while green is a symbol of Islam religion for the case of Senegal and fertility of the land for Mali. The green color stands for hope in the Ivory Coast flag and the catholic religion for Ireland, the orange color stands for the generous earth in the Ivory Coast, while in Ireland stands for the Protestants.

Table 2 Models' architecture

CNN Model	Model Type	Anchor based	Backbone	Neck	Head	Pre-trained Datasets
Fast-RCNN	Two-stage	Yes	VGG16	Roi Pooling (RP)	4 fully connected (FC) layers, classifier-softmax, bbox-regression-sigmoid	Pascal VOC 2012, ILSVRC
SSD	One-stage	Yes	VGG16	None	6 extra convolution layers with 3 × 3 kernel + sigmoid	Pascal VOC, MS COCO, ILSVRC
RetinaNet	One-stage	Yes	ResNet50	FPN	6 FC layers, classifier-softmax, bbox-regression-sigmoid	MS COCO
YoloV4	One-stage	Yes	CSPDarknet53	SPP+PAN	3 extra convolution layers for classification and regression with sigmoid	MS COCO
YoloV5	One-stage	Yes	CSPDarknet53	PAN	3 extra convolution layers for classification and regression with sigmoid	MS COCO
FCOS	One-stage	Noe	ResnetXt	FPN	2 FC layers, classifier-softmax, Centerness-softmax, bbox-regression-sigmoid	MS COCO
CornerNet	Two-stage	No	Hourglass	None	Residual blocks + Corner pooling layer + distance embedding module	MS COCO

Table 3 Models' optimizers' parameters and losses

Model name	Optimizer	Batch size	Learning rate	Epochs	Loss
Fast-RCNN	SGD	28	0.001	100	Multi-task loss (L1 for bbox + Softmax per ROI for class)
SSD	SGD	32	0.003	100	Weighted sum loss (Smooth L1 for bounding boc (bbox)+Softmax confidence for class)
RetinaNet	SGD	128	0.01	100	Focal loss
YoloV4	Adam	32	0.01	100	CIoU-loss (Focal loss for class, IOU for bbox)
YoloV5	Adam	32	0.01	100	CIoU-loss (Focal loss for class, IOU for bbox)
FCOS	SGD	16	0.0001	100	CIoU-loss (Focal loss for class, IOU for bbox)
CornerNet	Adam	49	0.0025	100	Variant Loss (detection loss, pull loss, push loss, offset loss)

Table 4 Computer specifications

CPU	GPU	RAM	Storage
AMD Ryzen Thread ripper 2920X (12-cores, 24-Threads)	NVIDIA GeForce RTX 2060 SUPER (8GB VRAM, 2176 cuda-cores)	32GB	M2-NVM 512 GB

Belgium and Germany have a different shine for their colors, with red color standing for victory, yellow for prosperity and black for humility in the case of Belgium. Germany flag's colors have different meanings, with red and gold being an inspiration from the Roman Empire army and black was formed after the movement against the Conservative European Order that was established after Napoleon's defeat and ensured the basic rights of German people.

India and Niger have common colors and designs but with different meanings also. India flag's green color stands for Muslim where in Niger stands for fertile regions of the country, the white color for India stands for peace among religions whereas for Niger stands for purity

Table 5 Tricolor flags feature inclusions

Flags pairs	YoloV4	YoloV5	Fast R-CNN	SSD	RetinaNet	CornerNet	FCOS	Flags Images
Chad - Romania	0.92	0.95	0.98	0.97	0.98	0.98	0.92	
Senegal - Mali	0.85	0.91	0.98	0.90	0.89	0.92	0.99	
Ireland - Côte d'Ivoire	0.89	0.87	0.94	0.93	0.64	0.96	0.95	
Belgium - Germany	0.91	0.93	0.98	0.99	0.97	0.98	0.98	
India - Niger	0.93	0.89	0.99	0.96	0.93	0.98	0.97	
Mexico - Italy	0.26	0.28	0.94	0.86	0.89	0.97	0.96	
Andorra - Moldova	0.95	0.88	0.96	0.99	0.97	0.98	0.99	
Bolivia - Ghana	0.96	0.93	0.96	0.95	0.98	0.98	0.99	
Tajikistan - Hungary	0.87	0.81	0.98	0.96	0.98	0.95	0.99	
Egypt - Iraq	0.98	0.91	0.97	0.98	0.98	0.99	0.99	
Netherlands - Paraguay	0.79	0.83	0.99	0.86	0.96	0.98	0.99	
Venezuela - Colombia	0.91	0.94	0.92	0.97	0.98	0.98	0.92	
Luxembourg - Netherlands	0.93	0.95	0.96	0.90	0.89	0.92	0.99	

and, finally, the orange color of Niger stands for the deserts whereas for India stands for courage and sacrifice.

Mexico’s and Italy’s flags share common meanings of colors where the green-white-red combination constitutes a symbol of republicanism, with red coming from bloodshed and white from peace among nations.

Venezuela and Colombia share common designs and colors. According to Vexillology, yellow stands for the riches of the countries, blue color stands for the sky and red for the bloodshed of the wars that the nations have met.

Luxembourg and Netherlands share common color meanings since Luxemburg’s flag adopted the colors from the Netherlands’s flag in a lighter version.

Tricolor flags have the most common historical, religious and political subjects of influence in the content of national flags. The most common meaning of red is bloodshed in all countries, while the white color stands for good values like morality and peace.

Table 6 includes the Tribar flags feature inclusion results. According to Table 6, Canada and Peru share common meanings where red stands for the wars and strength of the nations, and white color has the meaning of peace.

Despite Argentina’s and Nicargua’s flags have a common design and similar color tones, they have different meanings in their flags. In Argentina’s flag, the blue color stands for the sky and the white color stands for the clouds, while in Nicaragua’s flag, the blue color represents justice, and white represents peace.

Finally, in Austria’s and Latvia’s flags red and white colors stand for war and peace, respectively. Despite their different designs, tribar flags also share common characteristics like tricolors.

Table 7 includes the Bicolor flags feature inclusion results. The results are similar to the Tricolor stripe flags’ results, as the same symbolisms are shared.

For most of the examined pairs included in Tables 5, 6, and 7, the calculated feature inclusion for all seven object detection methods, is fully justified based on the Vexillology. The latter reveals that the proposed approach and inclusion measure is trustworthy and able to serve as a supporting tool to social sciences for the extraction of knowledge.

The same conclusions are drawn through the study of the results from the two remaining flag types.

Table 8 includes the Cross and Scandinavian Cross flags feature inclusion results. Both pairs in Table 8 have a cross, which is a common symbol despite their colors. The symbol is

Table 6 Tribar flags feature inclusions

Flags pairs	YoloV4	YoloV5	Fast R-CNN	SSD	RetinaNet	CornerNet	FCOS	Flags Images
Canada - Peru	0.93	0.89	0.92	0.99	0.97	0.98	0.98	
Argentina - Nicaragua	0.73	0.86	0.89	0.96	0.93	0.98	0.97	
Austria - Latvia	0.98	0.94	0.99	0.86	0.96	0.98	0.99	

Table 7 Bicolor flags feature inclusions

Flags pairs	YoloV4	YoloV5	Fast R-CNN	SSD	RetinaNet	CornerNet	FCOS	Flags Images
Indonesia - Monaco	0.98	0.95	0.94	0.97	0.98	0.98	0.92	
Haiti - Liechtenstein	0.67	0.72	0.99	0.90	0.89	0.92	0.99	
Indonesia - Poland	0.94	0.95	0.97	0.93	0.64	0.96	0.95	

Table 8 Cross and Scandinavian Cross flags feature inclusions

Flags pairs	YoloV4	YoloV5	Fast R-CNN	SSD	RetinaNet	CornerNet	FCOS	Flags Images
Dominica - Norway	0.88	0.86	0.8	0.97	0.98	0.98	0.92	
Dominica Republic - Iceland	0.96	0.93	0.61	0.86	0.96	0.98	0.99	

strongly correlated with Christianity religion, while Dominica and Dominica Republic were conquered by Nordic nations, which justifies the presence of their religions in their flags.

Table 9 includes the Cross and Scandinavian Cross flags feature inclusion results. The most dominant nation in canton flags is the Union Jack, which is the symbol of the United Kingdom, as well as a symbol for the United Kingdom Overseas Territories, which are 14 in total, with some of them presented in Table 6. These 14 countries have the Union Jack in their flag in the hoist area. Table 9 justifies the above fact, except for the last pair, referring to the feature inclusion of the United Kingdom in Australia, New Zealand, and Fiji.

The first pair New Zealand and Australia have also another common meaning despite the Union Jack, which is the stars in the flag that is a symbol for a constellation in the galaxy that can be seen only from their geographical location. Finally, United Stated of America (USA) and Liberia have different meanings in their flags' designs and symbols. Liberia's flag eleven stripes stand for the signatories of the Liberian Declaration of Independence where red and white colors symbolize courage and moral excellence, respectively. The USA national flag is designed based on navy flags, where the red and white stripes stand for the thirteen British colonies that declared independence from the Kingdom of Great Britain, and the 50 stars represent the states of the United States of America. Both Liberia and USA have common meanings in the design of their flags since independence played a significant role in their formation.

Closing the Vexillology-related analysis of the experimental results, it is clear that most correlated flags based on feature inclusion are also justified by Vexillology. Despite the differences between all the above presented national flags, there are common ideals and goals, like liberty and peace, encoded in their flags.

Combining Vexillology analysis and feature inclusion measures, it can be concluded that Tricolor, Tribar, and Bicolor flags interpret characteristics of a nation's history, while Canton flags interpret dominance or submission that can be confirmed based on history.

Results summarized in Tables 5, 6, 7, 8 and 9 also confirm the Hypotheses made. More specifically, Table 10 includes specific testing scenarios that verify the Hypotheses of this research.

Regarding Hypothesis #1, it can be assumed that certain flags are similar to each other, and the latter can be easily confirmed visually; results also indicate a high FI value for those flags, which additionally confirms numerically this assumption. Examples included in Table 10,

Table 9 Canton flags feature inclusions

Flags pairs	YoloV4	YoloV5	Fast R-CNN	SSD	RetinaNet	CornerNet	FCOS	Flags Images
New Zealand - Australia	0.74	0.78	0.98	0.97	0.98	0.98	0.92	
UK - New Zealand	0.36	0.37	0.28	0.31	0.27	0.33	0.24	
UK - Australia	0.38	0.34	0.26	0.31	0.27	0.33	0.24	
UK - Fiji	0.33	0.31	0.61	0.65	0.3	0.3	0.33	
USA - Liberia	0.89	0.88	0.98	0.86	0.96	0.98	0.99	

Table 10 Testing scenarios that verify the Hypotheses of this work

Hypothesis #1 *"There are flags of different nations that are similar to each other"*								
Flags pairs	**YoloV4**	**YoloV5**	**Fast R-CNN**	**SSD**	**RetinaNet**	**CornerNet**	**FCOS**	**Flags Images**
Egypt - Iraq	0.98	0.91	0.97	0.98	0.98	0.99	0.99	
Indonesia - Monaco	0.98	0.95	0.94	0.97	0.98	0.98	0.92	
Hypothesis #2 *"The proposed inclusion measure is able to measure the overall similarity between two flags"*								
Flags pairs	**YoloV4**	**YoloV5**	**Fast R-CNN**	**SSD**	**RetinaNet**	**CornerNet**	**FCOS**	**Flags Images**
Belgium - Germany	0.91	0.93	0.98	0.99	0.97	0.98	0.98	
Senegal - Liberia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Hypothesis #3 *"The proposed models are able to detect multiscale patterns into flags"*								
Flags pairs	**YoloV4**	**YoloV5**	**Fast R-CNN**	**SSD**	**RetinaNet**	**CornerNet**	**FCOS**	**Flags Images**
UK - New Zealand	0.36	0.37	0.28	0.31	0.27	0.33	0.24	
USA - Liberia	0.89	0.88	0.98	0.86	0.96	0.98	0.99	

indicate the cases of the flags of Egypt and Iraq and the flags of Indonesia and Monaco. As it can be observed, the pairs of flags are very similar to each other, and this is also evaluated numerically since FI values are high, between 0.91 and 0.99 for all models. Based on Vexillology, the FI is also historically matched; the Arab Liberation tricolor flag was inspired by the Egyptian Revolution and for this reason, Egypt's national flag forms the basis for the national flags of Egypt, Iraq, Sudan, Syria, and Yemen. The flags of Indonesia and Monaco are also identical. They both have two horizontal stripes, red over white. Their only difference is in their size; Indonesia's flag is slightly longer. FI calculation successfully quantified this similarity, returning values between 0.92 to 0.98. However, the Vexillology domain does not correlate the history of these two flags. Monaco's flag was based on the heraldic colors in the shield of the Monegasque princely arms, while the flag of Indonesia was influenced by its association with the Majapahit empire that used the red-and-white color combination flag for the Indonesian Navy. Based on Vexillology, these two flags embody different values, even though in both flags red color refers to flesh/blood and white to purity. Therefore, in this second testing scenario for Hypothesis #1, the flags are not historically connected; however, they display a high visual similarity, which is also confirmed and measured by the proposed FI value. It could be concluded that a high value of FI may raise questions about the connection of nations; if the nations are not historically connected, the above investigation may lead to connections that are not obvious, as in the above-mentioned case where the connection between Indonesia and Monaco lies solely in the interpretation of colors. The latter can establish the proposed method as an innovative tool for quantifying the connection between flags but also as a subjective criterion of similarity of flags that may raise questions for further investigation when FI value is high and the nations are not historically connected, thus strengthening the work of vexillologists and guiding their research.

Regarding Hypothesis #2, it can be assumed from the results that the proposed inclusion measure can provide a subjective measure of similarity, since its value is high for flags that are optically similar (e.g., Belgium-Germany) and low for flags that are different (e.g., Senegal-Liberia), as it can be seen in the testing scenarios included in Table 10.

Regarding Hypothesis #3, it can be verified that the proposed method is able to detect multiscale patterns into flags. For example, as it can be seen in Table 10, the flag of New Zealand shares the pattern of the UK flag. The latter is detected and quantified. In this case, FI values are low (between 0.24 and 0.37) since the similar pattern is subscaled, even though these nations are strongly related based on Vexillology, since New Zealand is a British colony. Therefore, it is verified once again that FI measure is an objective criterion of similarity, and it is able to measure the overall similarity of flags regardless of the nations' relations. In the case of USA and Liberia, the scaled pattern was also detected, i.e., red and white stripes, and therefore, since the strappy pattern is covering more of the flag area compared to the previous example, the FI value is higher (between 0.88 to 0.99) as expected.

The conducted testing scenarios are proof that the proposed methodology can be a useful tool for social science researchers since the similarity between flags can be objectively quantified for the first time with computational means by using optical patterns. When the proposed FI displays high values, connections between national flags can be verified by Vexillology; in cases where connections cannot be immediately verified and questions are raised, further investigation is required by experts to gain further insights.

5.3 Models performance and robustness

In this section, experimental results are evaluated regarding the models' performance. Performance results are presented for each model. Moreover, a comparison between the object detection models is conducted.

The proposed national flags dataset can be considered as a benchmark for the applied models since most of them have been tested in datasets with much fewer target classes than the proposed national flag dataset consisted of 195 classes, as it can be seen in Table 11.

In Table 11, only the SSD and the Fast R-CNN models have been applied in datasets with more target classes than the proposed nations' flag dataset. This fact gives the opportunity to analyze and benchmark the applied models from the perspective of robustness and sustainability in the huge information diversity that arises from the number of classes of the specific dataset. Therefore, in this section, the models are evaluated in terms of their behavior and robustness, to a problem that have never been applied. For this reason, models classification loss and models Bounding Box Regression Loss (BBox) over the training epochs are comparatively evaluated. These two performance plots are included in Fig. 5 for all examined models.

From Fig. 5, it can be seen that Fast R-CNN performed well despite the number of target classes, leading to the conclusion that the internal mechanisms of the model that generate and filter the huge amount of anchor boxes (region proposal) gave the potential to the object

Table 11 Selected models and their originally proposed datasets

Model Name	Dataset	Classes	Model Type
Fast R-CNN	Pascal VOC 2012	20	Two-stage, anchor-based
	ILSVRC	1000	
SSD	Pascal VOC	20	One-stage, anchor-based
	MS COCO	80	
	ILSVRC	1000	
RetinaNet	MS COCO	80	One-stage, anchor-based
YoloV4	MS COCO	80	One-stage, anchor-based
YoloV5	MS COCO	80	One-stage, anchor-based
FCOS	MS COCO	80	One-stage, anchor-free
CornerNet	MS COCO	80	Two-stage, anchor-free

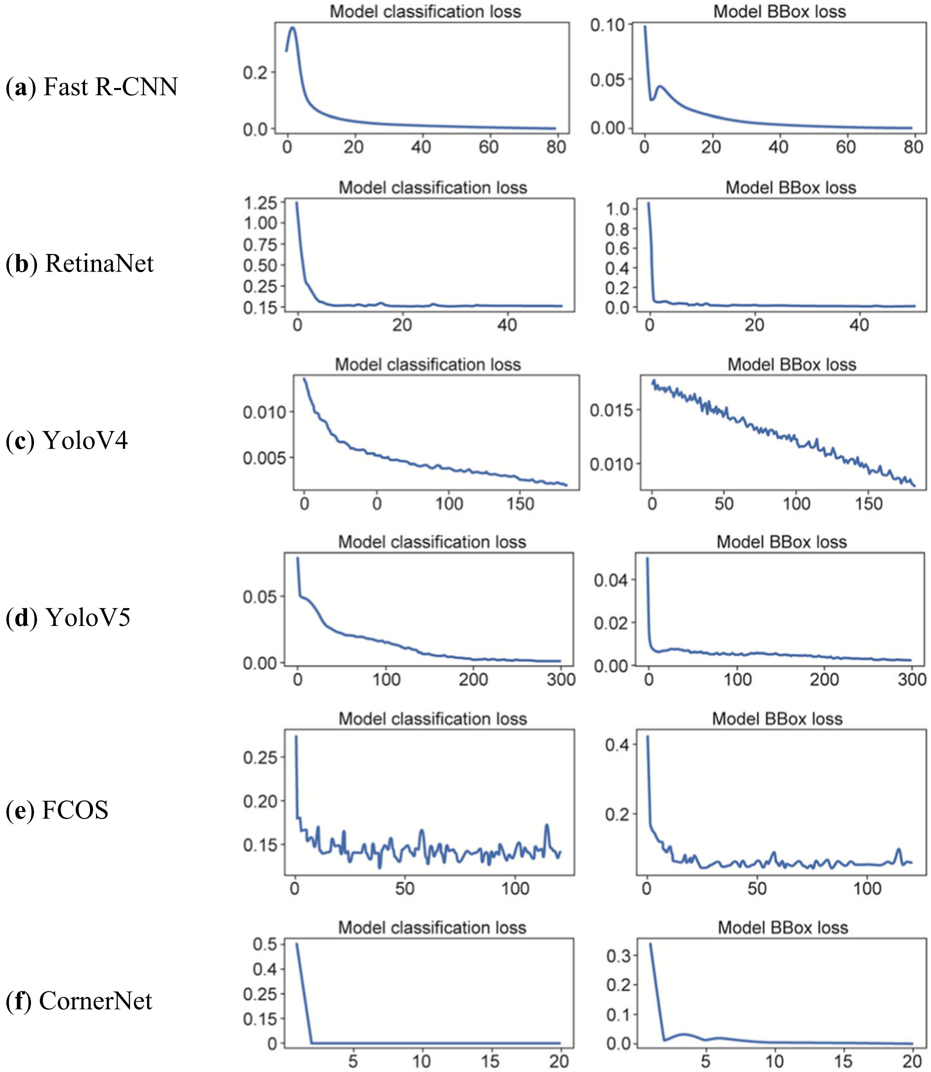

Fig. 5 Performance plots of classification loss and BBox regression loss during training: **a** Fast R-CNN; **b** RetinaNet; **c** YoloV4; **d** YoloV5; **e** FCOS; **f** CornerNet

detector to extract discriminative features for each target class. However, the regression part of the model did not perform equally with the classification. The latter is reasonable since the scales of the flags that were produced from augmentation and the rescaling from the backbone of the model created a redundancy, obstructing the model to generalize. Finally, the choice of the proposed regions (predictions) was based on a selection mechanism (based on higher classification confidence), which did not include a feature learning operation and could, therefore, lead to bad region selection.

The RetinaNet model also performed better in the classification part and worse in the regression part. In both cases, it performed better than Fast R-CNN. Based on the performance plots the mechanism in the loss function of RetinaNet is confirmed; it decreases the loss improvement in easy input data in order to give more effort in more hard cases. This characteristic is suitable for the specific problem since the most discriminative flags are easily manageable by the model and most of the training time the model tries to generalize in the more similar data and their target classes. Also, an advantage of the specific model in comparison to the previous one, Fast R-CNN, is that the feature maps produced from the ResNet backbone are fed to an FPN model to apply fusion in a diversity of feature map scales. This mechanism encapsulates a learning procedure, not exist in Fast R-CNN, which only applies a selection mechanism.

Regarding the performance of YoloV4 as illustrated in the performance plots of Fig. 5, it can be observed that the model achieved a steady decrease in the classification part while the regression part was slow to converge. Despite the difficulties of the dataset, YoloV4 was proven robust with high performance; yet the reported inclusion measure included in Tables 2 to 6 are among the lowest compared to the other models. The same can be observed for YoloV5; the model was slow to converge, especially in the classification part, as illustrated in Fig. 5. However, the classification performance was slightly better than that of the previous model, YoloV4. A general disadvantage of YoloV5, and the previous version YoloV4, is that discards useful information about object localization as the feature scale gets smaller in the architecture of the model. An advantage of this model for the specific problem is the stitching that applies in high- and low-resolution feature maps which leads to global features of the input image.

The FCOS model although it had fewer parameters than the other models, performed equally well. In addition, some strong characteristics were adopted from anchor-based one-stage object detectors, which were the multi-scale feature maps produced from the FPN model part. Therefore, the superiority of this new method in the calculation of bounding boxes' coordinates by FCOS over anchor-based models was justified by the resulted performance plots.

Finally, CornerNet performed better in classification loss than in BBox regression loss, which was at a similar level mostly at the final epochs. CornerNet could detect objects through keypoints that corresponded to the corners of the ground truth bounding box. As a general conclusion, it can be observed that CornerNet and RetinaNet, were the only models that performed similarly in both classification and BBox regression loss. Their performance plots revealed better behavior compared to other models since both appeared to converge quickly at a low value.

To summarize, five anchor-based and two anchor-free object detection models were evaluated for the national flag inclusion problem. Cumulatively, as it can be observed from Fig. 5, the anchor-free models performed better than anchor-based models; this can be attributed to the fact that anchor-free models control better the selection and discard of the proposed BBox points, which is a very important factor for the specific problem since the scale diversity among images is very large. Comparison between one-stage and two-stage object detectors, revealed similar performances, making it difficult to draw conclusions. Therefore, one-stage detectors overall may be a better choice than two-stage in the localization and classification task for the proposed dataset, since they come up with less computational demands. It can be noticed that all models are comparable and depending on the criteria we set each time, one may be better than another. In any case, what we are interested in is not to highlight the best model, but to study the behavior and robustness of all selected models to a problem that have never been applied.

6 Conclusions

This work aims to provide insights into the relationship between countries, through the lens of national flag similarity. Towards this end, a novel feature inclusion measure was proposed to quantify the correlations between national flags. The proposed measure was evaluated through the Vexillology domain and was proved to verify the obtained numerical results. The performance of six deep learning models (Yolo (V4 and V5), SSD, RetinaNet, Fast R-CNN, FCOS, and CornerNet) was investigated for the first time for the problem under study. The proposed models reported maximum feature inclusion between flags of up to 99%, quantifying and numerically justifying all similarities assumed by Vexillology. Results indicate that deep learning can be efficiently employed to quantitatively judge and numerically determine the correlation between national flags. Among the examined models, CornerNet and RetinaNet revealed better general behavior and a quicker convergence in both classification and BBox regression loss.

The proposed approach can be extended and applied to images of coats of arms and emblems of royal houses and empires to find correlations between them that could support and guide the historical study. Moreover, VQA (Visual Question Answering) methods could be applied to the above data source with the addition of Vexillology and history text to provide interpretability of historical artworks for didactic purposes. Towards this end, statistical methods and analysis tools could be implemented to support the presentation and teaching of nations' history through their flags and artworks, where these tools would be able to answer relevant questions. Another extension of the presented implementation could be the addition of the visual features that have changed in each nation's flag in time, in order to extract information about how one nation's historical events, e.g., independence or revolution, affected the colors or symbols of other national flags. Developing a specific customized model that could take into account the particularities of the problem under study could is also included in future work. Towards this direction, future work will focus on the integration of the proposed feature inclusion measure to the loss function to develop a model targeted to the problem. Finally, in future work, the proposed inclusion measure could be incorporated as a threshold for the anchor selection during training of the object detection model with national flag images to highlight the correlations of national flags during training rather than filtering the output.

Funding Open access funding provided by HEAL-Link Greece.

Data availability The dataset used in this study is available in https://github.com/MachineLearningVisionRG/MLVFlags.

Declarations

Consent for publication The authors of this work, consent to publish the work.

Conflict of interest There are no competing or conflicting interest to report for this work.

References

1. Akhand MAH, Hossain I, Murase K (2013) knowledge discovery from national flag through data mining approach. Int J Knowl Eng Res 2(4):212–216
2. Alzubi JA, Jain R, Nagrath P, Satapathy S, Taneja S, Gupta P (2021) Deep image captioning using an ensemble of CNN and LSTM based deep neural networks. J Intell Fuzzy Syst 40(4):5761–5769. https://doi.org/10.3233/JIFS-189415
3. Alzubi OA, Alzubi JA, Al-Zoubi AM, Hassonah MA, Kose U (2022) An efficient malware detection approach with feature weighting based on Harris hawks optimization. Clust Comput 25:2369–2387. https://doi.org/10.1007/s10586-021-03459-1
4. Assael Y, Sommerschield T, Prag J (2019) Restoring ancient text using deep learning: a case study on Greek epigraphy. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 6367–6374. https://doi.org/10.18653/v1/D19-1668
5. Bakkouri I, Afdel K (2020) Computer-aided diagnosis (CAD) system based on multi-layer feature fusion network for skin lesion recognition in dermoscopy images. Multimed Tools Appl 79(29–30):20483–20518. https://doi.org/10.1007/s11042-019-07988-1
6. Bakkouri I, Afdel K (2022) MLCA2F: multi-level context attentional feature fusion for COVID-19 lesion segmentation from CT scans. SIViP. https://doi.org/10.1007/s11760-022-02325-w
7. Belhi A, Bouras A, Foufou S (2018) Towards a hierarchical multitask classification framework for cultural heritage. 2018 IEEE/ACS 15th international conference on computer systems and applications (AICCSA), pp 1–7. https://doi.org/10.1109/AICCSA.2018.8612815
8. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) YOLOv4: optimal speed and accuracy of object detection. ArXiv. https://doi.org/10.48550/ARXIV.2004.10934
9. Bongini P, Becattini F, Bagdanov AD, Del Bimbo A (2020) Visual question answering for cultural heritage. IOP Conf Ser: Mater Sci Eng 949(1):012074. https://doi.org/10.1088/1757-899X/949/1/012074
10. Cerulo KA (1993) Symbols and the world system: national anthems and flags. Sociol Forum 8(2):243–271. https://doi.org/10.1007/BF01115492
11. Cilia ND, De Stefano C, Fontanella F, Marrocco C, Molinara M, Scotto Di Freca A (2020) An end-to-end deep learning system for medieval writer identification. Pattern Recogn Lett 129:137–143. https://doi.org/10.1016/j.patrec.2019.11.025
12. Gallwey J, Eyre M, Tonkins M, Coggan J (2019) Bringing lunar LiDAR Back down to earth: mapping our industrial heritage through deep transfer learning. Remote Sens 11(17):1994. https://doi.org/10.3390/rs11171994
13. Ghosh M, Obaidullah SM, Gherardini F, Zdimalova M (2021) Classification of geometric forms in mosaics using deep neural network. J Imaging 7(8):149. https://doi.org/10.3390/jimaging7080149
14. Girshick R (2015) Fast R-CNN. 2015 IEEE international conference on computer vision (ICCV), pp 1440–1448. https://doi.org/10.1109/ICCV.2015.169
15. Jboor NH, Belhi A, Al-Ali AK, Bouras A, Jaoua A (2019) Towards an Inpainting framework for visual cultural heritage. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), pp 602–607. https://doi.org/10.1109/JEEIT.2019.8717470
16. Jocher G, Chaurasia A, Stoken A, Borovec J, Kwon Y, Fang J, Michael K, Abhiram V, Minh DM, Nadar J, Skalski P, Wang Z, Hogan A, Fati C, Thanh LM, Patel D, Yiwei D, You F, Hajek J, Diaconu L (2022). Ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference (v6.1). Zenodo. https://doi.org/10.5281/zenodo.6222936
17. Kutlay MA, Yaman E (2016) Comparison of different machine learning algorithms for national flags classification. Southeast Eur J Soft Comput 4(2). https://doi.org/10.21533/scjournal.v4i2.94
18. Law H, Deng J (2020) CornerNet: detecting objects as paired Keypoints. Int J Comput Vis 128:642–656. https://doi.org/10.1007/s11263-019-01204-1
19. Lin T-Y, Goyal P, Girshick R, He K, Dollar P (2020) Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 42(2):318–327. https://doi.org/10.1109/TPAMI.2018.2858826

20. Lindauer MS (1969) Color preferences among the flags of the world. Percept Mot Skills 29(3):892–894. https://doi.org/10.2466/pms.1969.29.3.892
21. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2015) SSD: single shot MultiBox detector. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-319-46448-0_2
22. López-García P, Argote-Espino D, Fačevicová K (2018) Statistical processing of compositional data. The case of ceramic samples from the archaeological site of Xalasco, Tlaxcala, Mexico. J Archaeol Sci Rep 19: 100–114. https://doi.org/10.1016/j.jasrep.2018.02.023
23. Monisha GS, Malathi S (2021) Effective survey on handwriting character recognition. In: Advances in intelligent systems and computing, pp 115–131. https://doi.org/10.1007/978-981-15-7907-3_9
24. Movassagh AA, Alzubi JA, Gheisari M, Rahimi M, Mohan S, Abbasi AA, Nabipour N (2021) Artificial neural networks training algorithm integrating invasive weed optimization with differential evolutionary model. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-020-02623-6
25. Podeh E (2011) The symbolism of the Arab flag in modern Arab states: between commonality and uniqueness*. Nations Natl 17(2):419–442. https://doi.org/10.1111/j.1469-8129.2010.00475.x
26. Polak A, Kelman T, Murray P, Marshall S, Stothard DJM, Eastaugh N, Eastaugh F (2017) Hyperspectral imaging combined with data classification techniques as an aid for artwork authentication. J Cult Herit 26: 1–11. https://doi.org/10.1016/j.culher.2017.01.013
27. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 779–788. https://doi.org/10.1109/CVPR.2016.91
28. Sabatelli M, Kestemont M, Daelemans W, Geurts P (2019) Deep transfer learning for art classification problems, pp 631–646. https://doi.org/10.1007/978-3-030-11012-3_48
29. Sharafi S, Fouladvand S, Simpson I, Alvarez JAB (2016) Application of pattern recognition in detection of buried archaeological sites based on analysing environmental variables, Khorramabad plain, West Iran. J Archaeol Sci Rep 8:206–215. https://doi.org/10.1016/j.jasrep.2016.06.024
30. Sheng S, Moens M-F (2019) Generating captions for images of ancient artworks. Proceedings of the 27th ACM international conference on multimedia, pp 2478–2486. https://doi.org/10.1145/3343031.3350972
31. Tian Z, Shen C, Chen H, He T (2019) FCOS: fully convolutional one-stage object detection. Proceedings of the IEEE international conference on computer vision, pp 9626–9635. https://doi.org/10.1109/ICCV.2019.00972
32. Uehara Y, Endo S, Shiitani S, Masumoto D, Nagata S (2001) A computer-aided visual exploration system for knowledge discovery from images. Second international workshop on multimedia data mining (MDM/KDD’2001), pp 102–109
33. Wang X, Wang W, Li Z, Wang Y, Han Y, Hao Z (2018) A recognition method of the similarity character for Uchen script Tibetan historical document based on DNN. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), pp 52–62. https://doi.org/10.1007/978-3-030-03338-5_5

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.