This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2891745, IEEE Access

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Exploring Object-Centric and Scene-Centric CNN Features and their Complementarity for Human Rights Violations Recognition in Images

GRIGORIOS KALLIATAKIS[1], SHOAIB EHSAN[1], ALEŠ LEONARDIS[2], MARIA FASLI[1], AND KLAUS D. MCDONALD-MAIER.[1], (Senior Member, IEEE))

[1]School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K.
[2]School of Computer Science, University of Birmingham, Birmingham B15 2TT, U.K.

Corresponding author: Grigorios Kalliatakis (e-mail: gkallia@essex.ac.uk).

This work was supported by the Economic and Social Research Council (ESRC) under grant ES/ M010236/1, and Engineering and Physical Sciences Research Council (EPSRC) under grants EP/R02572X/1 and EP/P017487/1.

ABSTRACT Identifying potential abuses of human rights through imagery is a novel and challenging task in the field of computer vision, that will enable to expose human rights violations over large-scale data that may otherwise be impossible. While standard databases for object and scene categorisation contain hundreds of different classes, the largest available dataset of human rights violations contains only 4 classes. Here, we introduce the 'Human Rights Archive Database' (HRA), a verified-by-experts repository of 3050 human rights violations photographs, labelled with human rights semantic categories, comprising a list of the types of human rights abuses encountered at present. With the HRA dataset and a two-phase transfer learning scheme, we fine-tuned the state-of-the-art deep convolutional neural networks (CNNs) to provide human rights violations classification CNNs (HRA-CNNs). We also present extensive experiments refined to evaluate how well object-centric and scene-centric CNN features can be combined for the task of recognising human rights abuses. With this, we show that HRA database poses a challenge at a higher level for the well studied representation learning methods, and provide a benchmark in the task of human rights violations recognition in visual context. We expect this dataset can help to open up new horizons on creating systems able of recognising rich information about human rights violations.

INDEX TERMS Computer Vision; Image Interpretation; Visual Recognition; Convolutional Neural Networks; Human Rights Abuses Recognition

I. INTRODUCTION

Human rights violations have been unfolding during the entire human history, while nowadays they increasingly appear in many different forms around the world. By 'human rights violations' we refer in this paper to actions executed by state or non-state actors that breach any part of those rights which protect individuals and groups from behaviours which potentially interfere with fundamental freedoms and human dignity [1]. As mobile phones with photo and video capability are ubiquitous, individuals (human rights activists, journalists, eye witnesses and others) are recording and sharing high quality photos and videos of human rights incidents and circumstantial information. Photos and videos have become an important source of information for human rights investigations, including Commissions of Inquiry and Fact-finding Missions [36]. Investigators often receive digital images directly from witnesses, providing high quality corroboration of their testimonies. In most instances, investigators receive images from third parties (e.g. journalists or NGOs), but their provenance and authenticity is unknown as indicated in [2]. A third source of digital images is social media, e.g. uploaded to Facebook, again with uncertainty regarding authenticity or source [3]. Manually sifting through that sheer volume of images to verify if any abuse is taking place and then act on it, would be tedious and time-consuming work for humans. For this reason, a software tool aimed at identifying potential

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2891745, IEEE Access

TABLE 1. Proposed human rights violations categories with definitions.

1. Arms	Weapons systems that put civilians at high risk of armed conflict and violence
2. Child Labour	Work that deprives children of their childhood, their potential and their dignity, and that is harmful to physical and mental development
3. Child Marriage	A formal marriage or informal union before age 18. Child marriage is widespread and can lead to a lifetime of disadvantage and deprivation
4. Detention Centres	The right to health and a healthy environment, the right to be free from discrimination and arbitrary detention as critical means of achieving health
5. Disability Rights	People with disabilities experience a range of barriers to education, health care and other basic services, while they are subjected to violence and discrimination
6. Displaced Populations	Abuses against the rights of refugees, asylum seekers, and displaced people (block access to asylum, forcible return of people to places where their lives or freedom would be threatened, and deprive asylum seekers of rights to fair hearings of their refugee claims)
7. Environment	A lack of legal regulation and enforcement of industrial and artisanal mining, large-scale dams, deforestation, domestic water and sanitation systems, and heavily polluting industries can lead to host of human rights violations
8. Out of School	Discrimination of marginalized groups by teachers and other students, long distances to school, formal and informal school fees, and the absence of inclusive education are among the main causes of children staying out of school

abuses of human rights, capable of going through images quickly to narrow down the field would greatly assist human rights investigators.

The field of computer vision has developed several databases to organize knowledge about object categories [4]–[6], scenes [7]–[9] and materials [10]–[12]. However, an explicit image dataset of significant size depicting human rights violations does not currently exist. To our knowledge, the only attempt of constructing an image database in the context of human rights violations was presented in [13]. That dataset was limited to 4 different categories of human rights violations and 100 images per category, collected by utilising manually crafted query terms. Moreover, that dataset was assembled by images available on the Internet from unverified sources and does not offer high-coverage and high-diversity of exemplars.

In this paper, we describe in depth the construction of the *Human Rights Archive (HRA)* database, and evaluate the performance of several renowned convolutional neural networks for the task of recognising human rights violations, while acknowledging that our work is a first but significant step for human rights abuses analysis from single images. The objective of our work is not only to compare how features learned in object-centric CNNs and scene-centric CNNs perform, but also how they can complement each other when used as generic features in other visual recognition tasks. Also, we investigate the effects of different pooling strategies for efficient feature extraction and fusion. Finally, a visualisation of the important regions in the image for predicting target concepts, allows us to show differences in the internal representations of object-centric and scene-centric networks.

Major contributions in this work are as follows:

- A new, verified-by-experts dataset of human rights abuses, containing approximately 3k images for 8 violation categories, listed and defined in Table 1.
- We evaluate the performance of several state-of-the-art Convolutional Neural Networks (CNNs) for human rights violations recognition.
- We compare how the features learned in a CNN for scene classification (scene-centric) and features learned in a CNN for object classification (object-centric) behave when merged over different configurations.
- A web-demo for human rights violations recognition, accessible through computer or mobile device browsers.

The remainder of the paper is organized as follows. In Section II, we introduce the Human Rights Archive database, describe its unique collection procedure based on experts-verified sources, and present extended statistics. Section III delves into how transferable object-centric and scene-centric CNN features are for the task of classifying human rights violations, and introduces a web-demo for recognising human rights abuses in the wild from uploaded photos. Section IV investigates the complementarity of object-centric and scene-centric CNN features by exploiting different fusion mechanisms. The paper concludes in Section V.

II. HUMAN RIGHTS VIOLATIONS DATABASE

At present, organizations concerned with human rights advocacy are gradually using digital images as a tool for improving the exposure of human rights and international humanitarian law violations that may otherwise be impossible. However, in order to advance the automated recognition of human rights violations a well-sampled image database is required.

In this section, we describe the *Human Rights Archive (HRA)* database, a repository of approximately 3k well-sampled photographs of various human rights violations captured in real world situations and surroundings, labelled with 8 semantic categories, comprising the types of human rights abuses encountered around the world nowadays. Image samples are shown in Fig. 1. In order to increase the diversity of visual appearances in the HRA dataset (see Fig. 2), images from different situations or places are gathered. The dataset is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2891745, IEEE Access

FIGURE 1. Image samples from the categories of the Human Rights Archive (HRA) dataset. The dataset contains eight violation categories and a supplementary 'no violation' class.

FIGURE 2. Image samples from our human rights categories grouped by different situations to illustrate the diversity of the dataset. For each situation we show 3 labelled images.

available at https://github.com/GKalliatakis/Human-Rights-Archive-CNNs.

A. CHALLENGES

The fundamental asset of a high-quality dataset is a broad coverage of the explicit space that needs to be learned. The intention of Human Rights Archive database is to provide a collection of human rights violations categories encountered in the world nowadays, limited to activities that can be straightforwardly utilised to answer the question of whether there is a human right being violated in an image without any other prior knowledge regarding the action. To the best of our knowledge, the largest available dataset [13] in the context of human rights violations consists of only 4 classes and 100 images per category, with no other reference point in standardised dataset of images and annotations regarding human rights violations. The main drawback of that attempt is that the dataset was assembled by images available on the Internet from unverified sources and does not offer high-coverage and high-diversity of exemplars.

Human rights violations recognition is closely related to, but radically different from the tasks of object and scene recognition. As an example, one would easily correlate child labour with the task of recognising manual-labour-related tools (e.g., hoe and hammer). However, this would clearly be problematic for frequent cases such as adults working with those tools. The same applies for correlating a human right violation with the task of visual place recognition. For this reason, following a conventional image collection procedure is not appropriate for collecting images with respect to human rights violations. The first issue encountered is that the query terms for describing different categories of human rights violations must be provided by experts in the field of human rights and not by quasi-exhaustively searching a dictionary. The next obstacle concerns online search engines such as Google, Bing or even dedicated photo-sharing websites like Flickr, which returned a huge number of irrelevant results for the given queries of human rights violations and discussed in a previous study [14]. The final and most important matter of contention is the ground truth label verification of the images, which commonly is accomplished by crowd-sourcing the task to Amazon Mechanical Turk (AMT). However, in the case of human rights violations, human classification accuracy cannot be measured by utilising AMT for the reason that workers are not qualified for such specialised tasks.

B. BUILDING THE HUMAN RIGHTS ARCHIVE DATABASE

A key question with respect to the visual recognition problem of human rights violations from real-world images arises: how can this structured visual knowledge be gathered? As discussed in Section II-A, the crucial aspects of such a unique image database are the origin and the verification of the images. For this reason, and in order to obtain an adequate number of verified real-world images depicting human rights violations, we turn to non-governmental organizations (NGOs) and their public repositories.

The first NGO considered is Human Rights Watch which offers an online media platform (http://media.hrw.org/) capable of exposing human rights and international humanitarian law violations in the form of various media types such as videos, photo essays, satellite imagery and audio clips. Their online repository contains 9 main topics in the context of human rights violations (arms, business, children's rights, disabilities, health and human rights, international justice, LGBT, refugee rights and women rights) and 49 subcategories. In total, we download 99 available video clips from their online platform. After that, candidate images are being recorded for every video clip with a ratio of 10 (one image out of ten is recorded). This is done in order to obtain images distinctive enough on a frame to frame basis. Next, all the images that do not correspond to the definition of the human right violation category (mostly the interview parts of the clips) are manually removed. Images with low quality (very blurry or noisy, black-and-white), clearly manipulated (added text or borders, or computer-generated elements) or otherwise unusual (aerial views) are also removed. One considerable drawback in the course of that process is the presence of a watermark in most of the video files available from that platform. As a result, all the recorded images that originally contained the watermark had to be cropped in a suitable way. Only colour images of 600 x 900 pixels or larger were retrieved after the cropping stage. In addition to those images, all photo essays available for each topic and its subcategories are added, resulting in 342 more images to the final array. The entire pipeline used for collecting and filtering out the images from Human Rights Watch is depicted in Fig. 3.

The second NGO investigated is the United Na-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2891745, IEEE Access

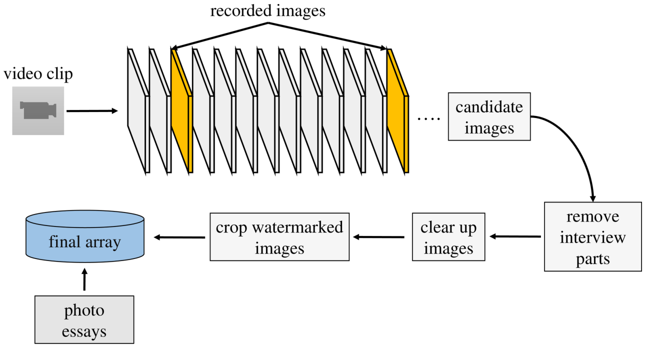

FIGURE 3. Image collection procedure from Human Rights Watch media repository.

FIGURE 4. t-SNE embedding of HRA dataset images based on their extracted features. Images that are nearby each other are also close in the CNN representation space, which implies that the CNN 'sees' them as being very similar. Notice that the similarities are more often class-based and semantic rather than pixel and colour-based.

tions which presents an online collection of images (http://www.unmultimedia.org/photo/) in the context of human rights. Their website is equipped with a search mechanism capable of returning relevant images for simple and complex query terms. In order to define a list of query terms, we utilise all main topics and their respective subcategories from Human Rights Watch and combine them with likely synonyms. For example, in order to acquire images depicting the employment of children in any work that deprives children of their childhood and interferes with their ability to attend regular school, 'child labour', 'child work' and 'child employment' were provided as queries to the database. In total, we download 8550 candidate images by utilising the list of query terms. We follow the same approach as Human Rights Watch in order to filter out the images. First, we manually remove all the images that do not correspond to the definition of the human right violation category. In the case of the United Nations online repository, the majority of the returned images showcased people sharing their testimony at various presentations or panel discussions. We also remove images that are black-and-white or otherwise unusual (aerial views). Finally, we add applicable high-resolution images to the database.

C. DATA ANALYSIS

In this section, we conduct in-depth analysis in various aspects of the dataset. The final dataset contains a set of 8 human rights violations categories and 2847 images (the number of images is continuously growing as we seek additional repositories verified by other NGOs), that cover a wide range of real-world situations. 367 ready-made images are downloaded from the two online repositories representing 12.88% of the entire dataset, while the remainder (2480) images are recorded from videos coming out of Human Rights Watch media platform. The categories are listed and defined in Table 1. Furthermore, 203 instances which are not considered as human rights violations, such as children playing and workers mining, have been incorporated into the database in order to assess the classification performance more precisely. Our human rights-centric dataset differs from the previous Human Rights UNderstanding (HRUN) dataset [13]. That dataset was created by collecting images available on the Internet using online search engines for different manually crafted terms, but the HRA database was created by collecting human rights violations categories from verified sources. Because some human rights violations are reported and documented more than others, the distribution of images is not uniform between the classes of the database, as seen in Table 2. Examples of human rights violations categories with more images are *child labour*, *displaced people*, and *environment*. Examples of under-sampled categories include *child marriage* and *detention centres*.

D. VISUALISING HRA

Convolutional neural networks can be interpreted as continuously transforming the images into a representation in which the classes are separable by a linear classifier. In order to obtain an estimation about the topology of the Human Rights Archive space, we examined the internal features learned by a CNN using t-SNE (t-distributed Stochastic Neighbour Embedding) [15] visualisation algorithm, by embedding images into two dimensions so that their low-dimensional representation has approximately equal distances as their high-dimensional representation. To produce that visualisation, we feed the HRA set of images through the well studied VGG-16 convolutional-layer CNN architecture [16], where the 4096 dimensional visual features are taken at the output of the second fully-connected layer (i.e., FC7) including the ReLU non-linearity by using caffe [17] framework. Those features are then plugged into t-SNE in order to project the image

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2891745, IEEE Access

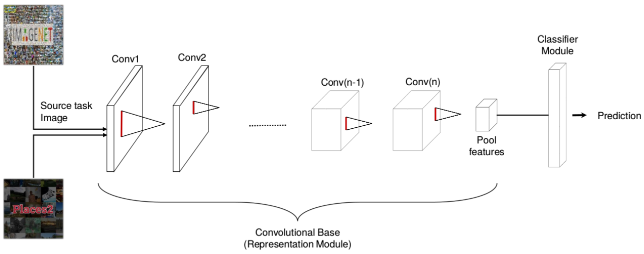

FIGURE 5. General structure of CNN architecture for end-to-end image classification.

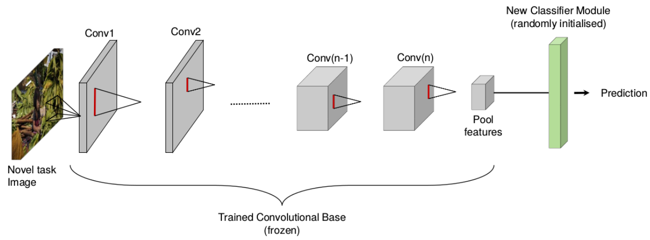

FIGURE 6. Network architecture used for high-level feature extraction with HRA. Pre-trained parameters of the internal layers of the networks are transferred to the target task. To compensate for the different nature of the source and target data we add a randomly initialised adaptation layer (fully connected layer) and train them on the labelled data of the target task.

features down to 2D. PCA preprocessing is used prior to the t-SNE routine to reduce to 10D to help optimize the t-SNE runtime. We then visualise the corresponding images in a grid as shown in Fig. 4, which can help us identify various clusters. Every position of the embedding is filled with its nearest neighbour. Note that since the actual embedding is roughly circular, this leads to a visualisation where the corners are a little 'stretched' out and over-represented.

III. CONVOLUTIONAL NEURAL NETWORKS FOR HUMAN RIGHTS VIOLATIONS CLASSIFICATION

A. IMPLEMENTATION

Given the impressive classification performance of the deep convolutional neural networks, we choose three popular object-centric CNN architectures, ResNet50 [18], VGG 16 convolutional-layer CNN [16], and VGG 19 convolutional-layer CNN [16], then fine-tune them on HRA to create baseline CNN models. Additionally, given the nature of the task at hand, we further fine-tuned a scene-centric CNN architecture, VGG16-Places365 [19] and compared it with the object-centric CNNs for human rights violations classification. We also trained a small CNN on the HRA training samples from scratch to set a baseline for what can be achieved. The baseline model is a simple stack of 3 convolution layers with a ReLU activation and followed by max-pooling layers. This is very similar to the architecture that LeCun et al. [20] advocated in the 1990s for image classification (with the exception of ReLU). Finally, we employed the above CNNs as fixed feature extractors by removing their classification block and computing a vector for every image in the HRA dataset, before training a nearest neighbour classifier with those extracted features. All the HRA-CNNs presented here were trained using the Keras package [21] on Nvidia GPU Tesla K80.

The baseline CNN contains 3.2 million parameters, while the other selected CNN architectures contain 138 million parameters for VGG16, 143 million parameters for VGG19 and 26 million parameters for ResNet50. VGG16-Places365 and VGG16 have exactly the same network architecture, but they are trained on scene-centric data and object-centric data respectively. Directly learning so many parameters from only a few thousand training images is problematic. A general structure of CNN architecture is depicted in Fig. 5.

B. TRANSFERRING CNN WEIGHTS

A conventional approach to enable training of very deep networks on relative small datasets is to use a model pre-trained on a very large dataset, and then use the CNN as as an initialization for the task of interest. This method, referred to as 'transfer learning' [22]–[24] injects knowledge from other tasks by deploying weights and parameters from a pre-trained network to the new one [25] and has become a commonly used method to learn task-specific features. The key idea is that the internal layers of the CNN can act as a generic extractor of high-level image representations, which can be pre-trained on one large dataset, the source task, and then re-used on other target tasks [26]. Considering the size of our dataset the chosen method to apply a deep CNN is to reduce the number of free parameters. In order to achieve this, the first filter stages can be trained in advance on different tasks of object or scene recognition and held fixed during training on human rights violations recognition. By freezing (preventing the weights from getting updated during training) the earlier layers, overfitting can be avoided. We initialize the feature extraction modules using pre-trained models from two different large scale datasets, ImageNet [27] and Places [19]. ImageNet is an object-centric dataset which contains images of generic objects including person and therefore is a good option for understanding the contents of the image region comprising the target person. On the contrary, Places is a scene-centric dataset specifically created for high level visual understanding tasks such as recognizing scene categories. Hence, pretraining the image feature extraction model using this dataset ensures providing global (high level) contextual support. For the target task (human rights violation recognition), we design a network that will output scores for the eight target categories of the HRA dataset or `no violation` if none of the categories are present in the image.

Feature extraction. Transfer is achieved in two phases. First, we start by using the representations learned by a previous network in order to extract interesting features from new samples. 'Feature extraction' consists of taking the convolutional base of a pre-trained network, running the new data of HRA through it and training a new, randomly initialised classifier on top of the semantic image output vector Y_{out}, as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2891745, IEEE Access

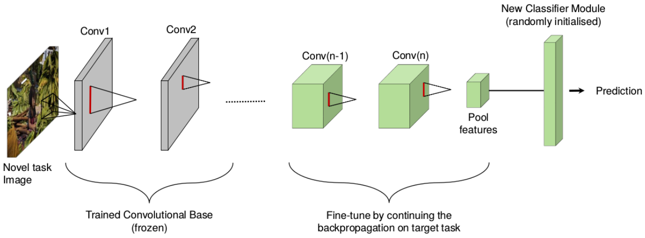

FIGURE 7. Network architecture used for fine-tuning with HRA. It marginally alters the more abstract representations of the model being utilised, in order to make them more relevant for the problem at hand.

illustrated in Fig.6. Note that Y_{out} is obtained as a complex non-linear function of potentially all input pixels and captures the high-level configurations of objects or scenes. We intentionally utilise only the convolutional base and not the densely-connected classifier of the original network, merely because the representations learned by the convolutional base are likely to be more generic. More importantly, representations found in densely-connected layers no longer contain any spatial information which is relevant for the task at hand. Note that in our experiments, the pooling layer just before the new classifier can be either a global average/max pooling operation for spatial data or simply a flattening layer. The FC_{HRA} layer compute $Y_{HRA} = \sigma(W_{HRA}Y_{out} + B_{HRA})$, where **W**, **B** are the trainable parameters. In all our experiments, the last convolutional layer of the pre-trained base have sizes (7, 7, 512).

Fine-tuning. The second phase required for transfer learning, complementary to feature extraction, is *fine-tuning*. Fine-tuning consists of unfreezing few of the top layers of a previously frozen convolutional base for feature extraction, and jointly training both the newly added fully-connected classifier and these top layers as illustrated in Fig. 7. It is only beneficial to fine-tune the top layers of the convolutional base once the classifier on top has already been trained (see Fig. 6). This is because the large gradient updates triggered by the randomly initialized weights would wreck the learned weights in the convolutional base. We choose to only fine-tune the last two convolutional layers rather than the entire network in order to prevent overfitting, since the entire network would have a very large entropic capacity and thus a strong tendency to overfit. The features learned by low-level convolutional layers are more general, less abstract than those found higher-up, so it is sensible to keep the first few layers fixed (more general features) and only fine-tune the last two (more specialized features).

For all of our experiments, we use the HRA dataset exclusively for the training process, while we obtain other representative images for each category from the Internet in order to compose the test set, producing a total of 270 reasonable images. Thus we eliminate the presence of bias in our experiments while our models are tested in the wild with real-world images. Table 2 summarises the statistics of the HRA dataset. For the purposes of our experiments,

TABLE 2. Statistics of the HRA dataset. The data is divided into two main subsets: training/validation data (trainval), and test data (test), with the trainval data further divided into suggested training(train) and validation (val) sets.

	train images	val images	trainval images	test images
arms	149	37	186	30
child labour	756	189	945	30
child marriage	69	18	87	30
detention centres	149	37	186	30
disability rights	218	55	273	30
displaced populations	487	122	609	30
environment	326	82	408	30
no violation	162	41	203	30
out of school	123	30	153	30
Total	2439	611	3050	270

TABLE 3. Classification accuracy and coverage on the test set of HRA for the deep features of various CNNs alongside two other baselines, a CNN trained from scratch (first row) and a nearest neighbour classifier for the extracted features (last four rows). Bold font highlights the dominant performance across the same metric.

	Pool	Top-1 acc.	Coverage	Train Params.
Baseline-CNN		12.59%	61%	3,240,553
VGG16	avg	34.44%	45%	4,853,257
VGG19	avg	**35.18%**	42%	4,853,257
ResNet50	avg	25.55%	55%	4,992,521
VGG16-places365	avg	30.00%	32%	4,853,257
VGG16	flatten	31.85%	55%	8,784,905
VGG19	flatten	31.11%	50%	8,784,905
ResNet50	flatten	30.00%	44%	4,992,521
VGG16-places365	flatten	28.51%	52%	8,784,905
VGG16	max	28.14%	**64%**	4,853,257
VGG19	max	29.62%	61%	4,853,257
ResNet50	max	25.55%	61%	4,992,521
VGG16-places365	max	26.66%	51%	4,853,257
VGG16 L2		22.59%	37%	-
VGG19 L2		24.44%	42%	-
ResNet50 L2		11.11%	18%	-
VGG16-places365 L2		18.51%	34%	-

the data is divided into two main subsets: training/validation data (trainval), and test data (test). To compensate for the imbalanced classes in HRA, we utilise cost-sensitive training to weight the loss function during training by an amount proportional to how under-represented each class is. This is useful to tell the model to 'pay more attention' to samples from an under-represented class. The maximum number of epochs was set to 40 iterations for each epoch and a learning rate of 0.0001, using the stochastic gradient descent (SGD) optimizer for cross-entropy minimization. The parameters were chosen empirically by analysing the training loss.

C. RESULTS ON HUMAN RIGHTS ARCHIVE (HRA)

After fine-tuning the various CNNs, we used the final output layer of each network to classify the test set images of HRA. In some applications it is possible for the system to refuse to make a decision. This is suitable when the algorithm can estimate how confident it should be about a decision, particularly if a wrong decision can be harmful and if a human operator is supposed to take over. Human rights violations recognition presents an example of this situation. Because the value of the recognition system deteriorates considerably if the prediction

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2891745, IEEE Access

TABLE 4. Classification accuracy and coverage on the test set of HRA without weighting the loss function during training. Bold font highlights the dominant performance across the same metric.

	Pool	Top-1 acc.	Coverage	Train Params.
Baseline-CNN		15.55%	34%	3,240,553
VGG16	avg	25.92%	23%	4,853,257
VGG19		24.07%	32%	4,853,257
ResNet50		17.40%	2%	4,992,521
VGG16-places365		26.66%	16%	4,853,257
VGG16	flatten	27.40%	41%	8,784,905
VGG19		28.88%	41%	8,784,905
ResNet50		18.50%	4%	4,992,521
VGG16-places365		25.55%	49%	8,784,905
VGG16	max	**28.51%**	38%	4,853,257
VGG19		22.22%	**53%**	4,853,257
ResNet50		10.74%	2%	4,992,521
VGG16-places365		25.55%	40%	4,853,257

TABLE 5. Classification accuracy and coverage on the test set of HRA using real-time data augmentation during training. Bold font highlights the dominant performance across the same metric.

	Pool	Top-1 acc.	Coverage	Trainable Params.
Baseline-CNN		13.70%	17%	3,240,553
VGG16	avg	33.70%	37%	4,853,257
VGG19		32.59%	33%	4,853,257
ResNet50		24.81%	59%	4,992,521
VGG16-places365		25.18%	26%	4,853,257
VGG16	flatten	**34.07%**	34%	8,784,905
VGG19		**34.07%**	27%	8,784,905
ResNet50		24.44%	64%	4,992,521
VGG16-places365		26.29%	37%	8,784,905
VGG16	max	32.22%	43%	4,853,257
VGG19		27.40%	**60%**	4,853,257
ResNet50		22.96%	54%	4,992,521
VGG16-places365		24.07%	43%	4,853,257

for meaningful images is inaccurate, it is important to point out images that depict human rights violations only if the confidence of the prediction is above a threshold. Of course, an automated system is only useful if it is able to effectively reduce the amount of photos that a human rights investigator must process. A realistic performance metric to use in this situation is *coverage*. Coverage is the proportion of a data set for which a classifier is able to produce a prediction. The classification results, using the cost-sensitive training, for *top-1 accuracy* and *coverage* are listed in Table 3. For the sake of completeness, we also provide classification results without weighting the loss function during training as illustrated in Table 4 and with real time data augmentation during training in Table 5. These rather weak scores suggest that our initial intuition of training imbalanced classes equitably by increasing the importance of the under-represented classes has indeed a positive effect on both accuracy and coverage. Although on paper applying a number of random transformations in order to augment our training samples will help the models generalise better, as revealed by similarly low scores, data augmentation does not improve the accuracy and coverage of the models for most of the cases. Note that for all the remaining experiments presented in this paper, results concerning only the superior cost-sensitive training are indicated. Given that a system capable of recognising human rights violations from visual content is only useful if they have high coverage, it was important to set a high coverage requirement for this task. Specifically, the network refuses to classify an input x, whenever the probability of the output sequence $p(y|x) < t$ for some confidence threshold t. For all the experiments in this paper, we set the confidence threshold at 0.85 in order to report the coverage performance.

Fig. 8 shows the responses to examples predicted by the best performing HRA-CNN, VGG19. Broadly, we can identify one type of misclassification given the current label attribution of HRA: images depicting the evidence which are responsible for a particular situation and not the actual action, such as schools being targeted by armed attacks. Future development of the HRA database, will explore to assign multi-ground truth labels or free-form sentences to images to better capture the richness of visual descriptions of human rights violations.

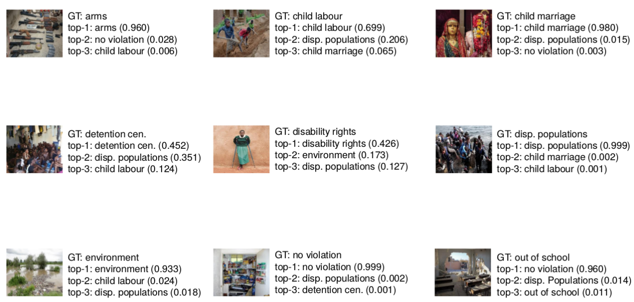

FIGURE 8. The predictions given by the best performing HRA-VGG19 for the images from the test set. The ground-truth label (GT) and the top 3 predictions are shown. The number beside each label indicates the prediction confidence.

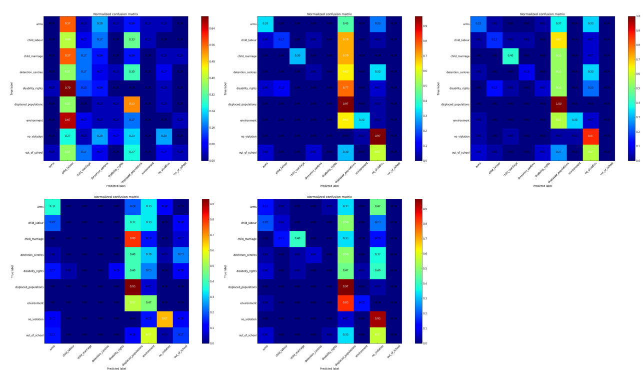

FIGURE 9. Normalized confusion matrices of the best performing CNNs: (left to right) first row: Baseline-CNN, VGG16-avg, and VGG19-avg; second row: ResNet50-flatten, and VGG16-places365-avg.

Fig. 9 illustrates the normalised confusion matrices of the best performing CNNs. These results indicate that predictions relying solely on object-based information are likely to misinterpret visual samples that belong to the class of disability rights as displaced populations. Other examples where the CNNs make mistakes are: predicting detention centres as displaced populations, out of school as no violation. This is not surprising because these pairs share similar properties,

FIGURE 10. Given an input image A, we visualise the class-discriminative regions of different CNNs using Grad-CAM [29] for the output class 'displaced populations'. The object-centric model B focuses on the head of the people, while the scene-centric model C focuses on the shelters in the background.

e.g., numerous people gathered at one place.

We can see that both VGG architectures surpass the scene-centric architecture of VGG16-Places365 by a significant margin of at least 4.44% for top-1 accuracy and 10% for coverage for their best performing pooling operation, even though the number of trainable parameters remains exactly the same. On the other hand, VGG16-Places365 outperform the object-centric ResNet50 for two of the pooling schemes. We have also tried to change the number of layers which were fine-tuned in our training set-up. Increasing the number of layers to three results in about 7% drop in classification performance. It is evident from Table 3 that each object-centric and scene-centric CNN has different strengths and weaknesses. Therefore, we expect that using an ensemble of different models would further boost the accuracy for the task of recognising human rights violations.

D. MODEL INTERPRETATION

In order to interpret which parts of a given image led a CNN to its final prediction, we produce heatmaps of 'class activation'. Class Activation Mapping (CAM) [28] and its generalisation Gradient-weighted Class Activation Mapping (Grad-CAM) [29] visualise the linear activations of a late layer's activations with respect to the class considered. To generate Grad-CAM visual explanations, we followed the approach of [29]. An image is fed into the fine-tuned network and the output feature maps of the last convolutional layer are extracted. Convolutional features are capable of retaining spatial information compared to fully-connected layers where that information is lost. The gradient of the score associated with a specific output class is computed, with respect to the extracted feature maps of the last convolutional layer. Then, the gradients are global-average-pooled to obtain the importance weights. Finally, the Grad-CAM is obtained by performing a weighted combination of forward activation maps followed by a ReLU. Fig. 10 shows an example of Grad-CAMs for the output class of 'displaced populations'.

E. WEB-BASED SOFTWARE FOR HUMAN RIGHTS VIOLATIONS RECOGNITION

Based on our trained HRA-CNNs, we created a web-demo for human rights violations recognition[1], accessible through computer or mobile device browsers. It is possible to upload photos to the web-based software to identify if images depict a human right violation, while the system suggests the 3 most likely semantic categories from the HRA dataset. A screenshot of the prediction result on a web browser is shown in Fig. 11. More precisely, the Keras [21] python deep learning framework over TensorFlow [30] was used to train the back-end prediction model in the demo. With this system, those combating abuse will be able to go through images very quickly to narrow down the field and identify pictures which need to be looked at in more detail. Furthermore, with the extensive use of this software, we will collect an expanded range of images depicting human rights abuses, in order to enhance the accuracy of our CNN models with larger data sets. Future directions for this work will include the capacity to receive feedback from people regarding the result. The source code for the web-based software is available at https://github.com/GKalliatakis/Human-Rights-Archive-CNNs to assist future research.

[1] http://83.212.117.19:5000/

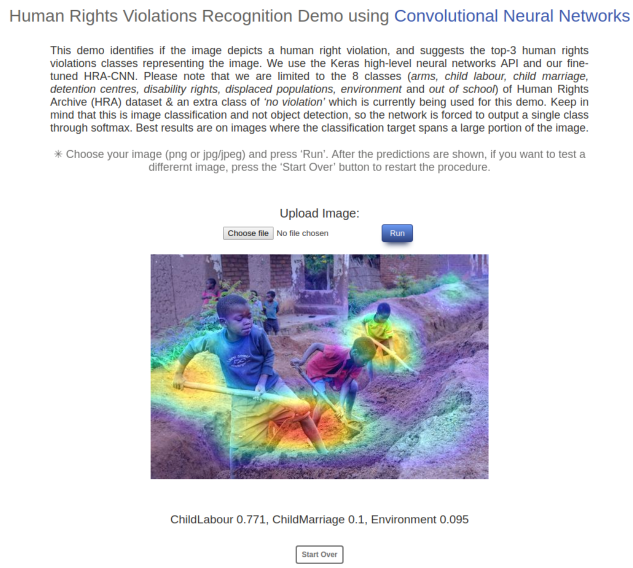

FIGURE 11. A screenshot of the human rights violations recognition demo based on the fine-tuned HRA-CNN. The web-demo predicts the type of human right that is being violated for uploaded photos.

IV. COMBINING SCENE-CENTRIC AND OBJECT-CENTRIC CNNS

Information fusion can be a crucial component in image classification schemes where increasing the overall accuracy of the system is regarded as one of its most integral aspects. Merging different information is not only meaningful because of the accuracy improvement it might offer in a system, but also for allowing the system to be more robust against changing dynamics. Since scenes are composed in part of objects, accurate recognition of human rights violations requires knowledge about both scenes and objects. By visualising the class-discriminative regions of object-centric-CNNs and scene-centric-CNNs (see Fig. 10) we find that both focus on different aspects of the image in order to classify it. Inspired by that observation we want to investi-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2891745, IEEE Access

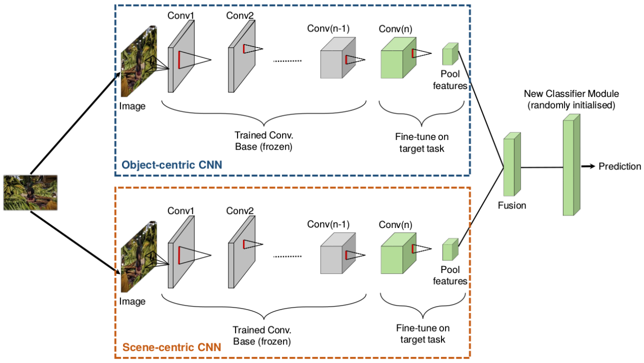

FIGURE 12. Early fusion.

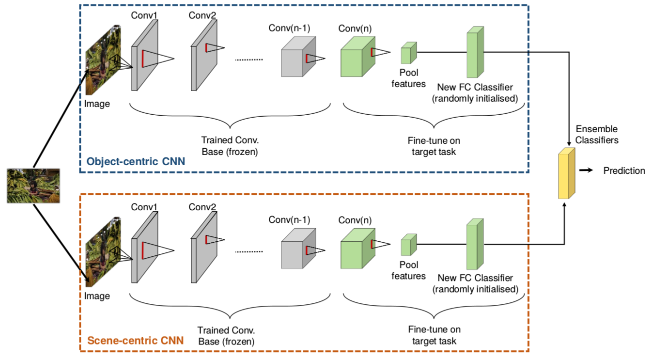

FIGURE 13. Late fusion.

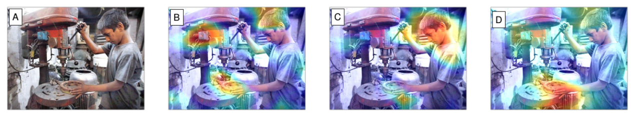

FIGURE 14. Informative region for predicting the category 'child labour' for different CNNs. Given an input image A, we visualise the class-discriminative regions using Grad-CAM [29] for the output class. The object-centric model B focuses on the tool used by the young boy, the scene-centric model C focuses mostly on the head of the young boy, while the early fusion of the two CNNs D focuses more on what the young boy is holding.

gate whether feature fusion (accommodating features coming from different sources into a single representation), which has resulted in increased performances in recent works [32], [33], would have similar effects for the unproven task of recognising human rights violations.

As indicated by [34], [35], feature fusion approaches can be grouped into two main categories: *early* and *late* fusion. We employ early and late fusion schemes in different ways along with the CNN architectures. The processing pipelines of our early and late fusion schemes are depicted in Fig. 12 and Fig. 13 respectively.

A. PROPOSED FUSION SCHEMES

1) Early fusion

Suppose we are given two CNNs, an object-centric network and a scene-centric network. Let feature set $F = \{f_1, f_2\}$ be extracted features of the last convolutional layer from each network, where each feature is a high-dimensional feature vector represented with $f_i \in \mathbb{R}^{d_i}$. Every distinct feature may have different cardinality according to the particular CNN architecture, such that $d_i = \{d_1, d_2\}$. Then the feature fusion function ϕ can be defined as the mapping operator on F such that $\phi(F) \mapsto R^d$.

The first strategy exploited in the early fusion scheme is the *concatenation* method, where discrete feature vectors of different sources are concatenated into one super-vector $f_f = \{f_1, f_2\}$ which will represent the final image feature. The final vector size is the summation of all feature dimensions $d = \sum_{i=1}^{n} d_i$.

The second fusion strategy employed is *averaging*, also known as *sum pooling* in the context of neural networks. In this strategy, the feature set F is averaged in order to form the final image descriptor $f_f = \frac{1}{n}\sum_{i=1}^{n} f_i$. All features in F should either have the same cardinality or the feature dimensions must be normalised prior to the fusion operation.

The last fusion strategy utilised is *maximum pooling*. It involves the same preprocessing in terms of the final feature cardinality, however it varies in the way features are merged. Maximum pooling selects the highest value from the corresponding features instead of taking the average of all features elements' in sum pooling. If the final feature representation is $f_f \mapsto R^d$, then max pooling selects each member of f_f as $f_f^i = \max_{i=1}^{d}(f_1^i, f_2^i)$.

2) Late fusion

Contrary to early fusion, the late fusion scheme consists of pooling together the predictions of a set of different end-to-end models (in our case object-centric and scene-centric CNNs), to produce more accurate predictions. This kind of assemblage relies on the assumption that independently trained object-centric and scene-centric models are focusing on slightly different aspects of the data to make their predictions as illustrated in Fig. 10. The easiest way to pool the predictions of a set of classifiers is to average their predictions at inference time as illustrated by Fig. 13.

B. DIFFERENCES AND COMPLEMENTARITIES

After evaluating deep features of various object-centric and scene-centric CNNs on the test set of HRA, we turn our attention to the problem of combining those features for the same task. First, we start by transferring CNN weights as described previously, this time combining the output of the last convolutional layer of an object-centric CNN with the output of a scene-centric CNN before randomly initialising a new fully-connected classifier as shown in Fig. 12. Note that in this approach only the last convolutional layer of each network is fine-tuned in order to keep equal number of trainable parameters with the previous set-up. We compare results of three fusion and pooling operations and their combinations as illustrated in Table 6.

Remarkably, results indicate that early fusion of object-centric and scene-centric features constantly trail their in-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2891745, IEEE Access

TABLE 6. Classification accuracy/precision and coverage on the test set of HRA using early fusion of deep features. Bold font highlights the dominant performance across the same metric.

	Pool	Fusion	Top-1 acc.	Coverage	Train Params.
VGG16 + VGG16-places365	avg	average	31.48%	14%	4,853,257
		concatenate	30.37%	25%	4,984,329
		maximum	30.74%	19%	4,853,257
	flatten	average	27.40%	57%	11,144,713
		concatenate	27.77%	58%	17,567,241
		maximum	29.25%	54%	11,144,713
	max	average	25.18%	45%	4,853,257
		concatenate	27.40%	49%	4,984,329
		maximum	24.44%	56%	4,853,257
VGG19 + VGG16-places365	avg	average	27.03%	14%	4,853,257
		concatenate	27.77%	25%	4,984,329
		maximum	28.88%	28%	4,853,257
	flatten	average	25.55%	**64%**	11,144,713
		concatenate	28.88%	50%	17,567,241
		maximum	28.14%	51%	11,144,713
	max	average	27.03%	37%	4,853,257
		concatenate	26.29%	47%	4,984,329
		maximum	26.29%	48%	4,853,257
ResNet50 + VGG16-places365	avg	average	27.03%	5%	2,494,473
		concatenate	28.51%	14%	2,625,545
		maximum	**31.85%**	16%	2,494,473
	flatten	average	27.77%	41%	8,785,929
		concatenate	27.03%	50%	15,208,457
		maximum	24.81%	50%	8,785,929
	max	average	25.92%	34%	2,494,473
		concatenate	25.55%	41%	2,625,545
		maximum	31.11%	30%	2,494,473

dividual counterparts in most of the evaluations for both performance metrics. More precisely, the global best coverage of 64% (VGG19 with average pooling) can be matched by *VGG19+VGG16-places365* when features were flattened and average fusion was utilised. However, the global best accuracy of 35.18% surpasses all early fusion schemes by a significant margin of at least 3.33%. An interesting observation is that object-centric features complement effectively the accuracy of their scene-centric counterparts in numerous combinations. The overall best results for early fusion are achieved when combining ResNet50 with VGG16-places365. Negative effects seem to occur mostly when combining very similar models like VGG16, VGG19 and VGG16-places365 which are all based on the same architecture. As opposed to [32], [33] where different data fusion strategies significantly increase performance, the accuracy does not see marked improvement in the case of human rights violations. Through the visualisation of the class-discriminative regions in Fig. 14, we can have a better understanding of what has been learned inside the CNNs for the early fusion scheme.

Regarding late fusion, although the ensemble classifier of object-centric and scene-centric models fall behind their individual counterparts in most of the evaluations, combining classifier pairs significantly improves the results compared to early fusion, making them almost all acceptable both in accuracy and coverage as shown in Table 7.

TABLE 7. Classification accuracy/precision and coverage on the test set of HRA using late fusion (ensemble the classifiers). Bold font highlights the dominant performance across the same metric.

	Pool	Accuracy	Coverage
		Top-1	Fraction
VGG16+VGG16-places365	avg	**32.92%**	31%
VGG19+VGG16-places365	avg	32.22%	29%
ResNet50+VGG16-places365	avg	28.88%	25%
VGG16+VGG16-places365	flatten	28.88%	42%
VGG19+VGG16-places365	flatten	28.14%	38%
ResNet50+VGG16-places365	flatten	28.88%	26%
VGG16+VGG16-places365	max	28.51%	41%
VGG19+VGG16-places365	max	27.40%	**44%**
ResNet50+VGG16-places365	max	28.51%	33%

V. CONCLUSION

This paper investigates the problem of human rights violations prediction from a single image. We present the HRA database, a dataset of images in non-controlled environments containing activities which reveal a human right being violated without any other prior knowledge. The images derive from experts-verified repositories and are labelled with 8 violations categories, proposed and described in this work. Using this dataset and a two-phase deep transfer learning scheme, we conduct an evaluation of recent deep learning algorithms and provide a benchmark on the proposed problem of visual human rights violations recognition. We also presented a thorough investigation on the relevance of combining object-centric and scene-centric CNN features alongside their differences and complementaries. These results reinforce the view that although human rights violations recognition is closely related with the simple tasks of object

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2891745, IEEE Access

and scene recognition, it poses a challenge at a higher level for the well studied representation learning methods. A technology capable of identifying potential human rights abuses in the same way as humans do has a lot of potential applications in human-assistive technologies and would greatly support human rights investigators.

REFERENCES

[1] Publication UN. Training Manual on Human Rights Monitoring; 2001. Available from: http://www.ohchr.org/EN/PublicationsResources/Pages/MethodologicalMaterials.aspx.

[2] Matheson K. Video as Evidence Field Guide;. Available from: https://witness.org/.

[3] McPherson E. Advocacy OrganizationsâĂŹ Evaluation of Social Media Information for NGO Journalism: The Evidence and Engagement Models. American Behavioral Scientist. 2015;59(1):124–148.

[4] Griffin G, Holub A, Perona P. Caltech-256 Object Category Dataset. California Institute of Technology; 2007. 7694.

[5] Torralba A, Fergus R, Freeman WT. 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2008;30(11):1958–1970.

[6] Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding. 2007;106(1):59–70.

[7] Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. Learning deep features for scene recognition using places database. In: Advances in neural information processing systems; 2014. p. 487–495.

[8] Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A. Sun database: Large-scale scene recognition from abbey to zoo. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE conference on. IEEE; 2010. p. 3485–3492.

[9] Zhou B, Khosla A, Lapedriza A, Torralba A, Oliva A. Places: An image database for deep scene understanding. arXiv preprint arXiv:161002055. 2016;.

[10] Liu C, Sharan L, Adelson EH, Rosenholtz R. Exploring features in a bayesian framework for material recognition. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE; 2010. p. 239–246.

[11] Sharan L, Rosenholtz R, Adelson E. Material perception: What can you see in a brief glance? Journal of Vision. 2009;9(8):784–784.

[12] Bell S, Upchurch P, Snavely N, Bala K. Material recognition in the wild with the materials in context database. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition; 2015. p. 3479–3487.

[13] Kalliatakis G, Ehsan S, Fasli M, Leonardis A, Gall J, McDonald-Maier KD. Detection of Human Rights Violations in Images: Can Convolutional Neural Networks Help? In: Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP, (VISIGRAPP 2017). INSTICC. SciTePress; 2017. p. 289–296.

[14] Kalliatakis G, Ehsan S, McDonald-Maier KD. A Paradigm Shift: Detecting Human Rights Violations Through Web Images. In: Proceedings of the Human Rights Practice in the Digital Age Workshop; 2017.

[15] van der Maaten L, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research. 2008;9:2579–2605.

[16] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR. 2014;abs/1409.1556.

[17] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. In: Proceedings of the 22Nd ACM International Conference on Multimedia. MM ’14. ACM; 2014. p. 675–678.

[18] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770–778.

[19] Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;.

[20] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition Proceedings of the IEEE, 86(11), p. 2278-2324.

[21] Chollet F, et al.. Keras; 2015. https://github.com/fchollet/keras.

[22] Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering. 2010;22(10):1345–1359.

[23] Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, et al. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In: ICML; 2014. p. 647–655.

[24] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Springer; 2014. p. 818–833.

[25] Kalliatakis G, Stamatiadis G, Ehsan S, Leonardis A, Gall J, Sticlaru A, et al. Evaluating Deep Convolutional Neural Networks for Material Classification. In: Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP, (VISIGRAPP 2017). INSTICC. SciTePress; 2017. p. 346–352.

[26] Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014. p. 1717–1724.

[27] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. NIPS’12. Curran Associates Inc.; 2012. p. 1097–1105.

[28] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 2921–2929.

[29] Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. arXiv preprint arXiv:161002391. 2016;.

[30] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467. 2016;.

[31] Mangai UG, Samanta S, Das S, Chowdhury PR. A survey of decision fusion and feature fusion strategies for pattern classification. IETE Technical review. 2010;27(4):293–307.

[32] Wang L, Wang Z, Du W, Qiao Y. Object-scene convolutional neural networks for event recognition in images. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops 2015 p. 30-35.

[33] Wu Z, Fu Y, Jiang YG, Sigal L. Harnessing object and scene semantics for large-scale video understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016 p. 3112-3121.

[34] Snoek CG, Worring M, Smeulders AW. Early versus late fusion in semantic video analysis. In: Proceedings of the 13th annual ACM international conference on Multimedia. ACM; 2005. p. 399–402.

[35] Ergun H, Akyuz YC, Sert M, Liu J. Early and Late Level Fusion of Deep Convolutional Neural Networks for Visual Concept Recognition. International Journal of Semantic Computing. 2016;10(03):379–397.

[36] Alston P, Knuckey S. The transformation of human rights fact-finding IOxford University Press. 2016;

GRIGORIOS KALLIATAKIS received the B.Sc. and M.Sc. degrees in informatics engineering and informatics & multimedia from Technological Education Institute (TEI) of Crete, Greece, in 2012 and 2015, respectively, and a M.Sc. degree in computer vision from University of Burgundy, France, in 2015. He has been pursuing the Ph.D. in Computer Science from the University of Essex, UK, since 2015. He is currently working on visual recognition of human rights abuses. His research interests include computer vision, image interpretation and visual recognition.

SHOAIB EHSAN received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2003, and the Ph.D. degree in computing and electronic systems (with specialization in computer vision) from the University of Essex, Colchester, U.K., in 2012. He has an extensive industrial and academic experience in the areas of embedded systems, embedded software design, computer vision, and image processing. His current research interests are in intrusion detection for embedded systems, local feature detection and description techniques, and image feature matching and performance analysis of vision systems. He was a recipient of the University of Essex Post Graduate Research Scholarship and the Overseas Research Student Scholarship and also a recipient of the prestigious Sullivan Doctoral Thesis Prize awarded annually by the British Machine Vision Association.

ALEŠ LEONARDIS is currently an Adjunct Professor with the Faculty of Computer Science, Graz University of Technology. His research interests include robust and adaptive methods for computer vision, object and scene recognition and categorization, statistical visual learning, 3-D object modeling, and biologically motivated vision. He is currently a Chair of Robotics with the University of Birmingham.

MARIA FASLI is currently a Professor of Computer Science (Artificial Intelligence) and the Director of the Institute for Analytics and Data Science (IADS) at the University of Essex. She obtained her BSc in Informatics from the Technological Education Institute of Thessaloniki in 1996, and her PhD in Computer Science from the University of Essex in 2000. She has held research and academic positions at the University of Essex since 1999 and became Professor in 2012. In 2009, she became the Head of the School of Computer Science and Electronic Engineering at Essex, a post which she held until the end of 2014. In August 2014, she was appointed in her current role as Director of IADS. In 2016, she was awarded the first UNESCO Chair in Analytics and Data Science. Her research interests lie in artificial intelligence techniques for analyzing and modeling complex systems and structured and unstructured data in various domains. Her research has been funded by National Research Councils in the UK, InnovateUK as well as businesses.

KLAUS D. MCDONALD-MAIER (S'91-SM'06) is currently the Head of the Embedded and Intelligent Systems Laboratory, University of Essex, Colchester, U.K. He is also Chief Scientist of UltraSoC Technologies Ltd., CEO of Metrarc Ltd., and a Visiting Professor with the University of Kent. His current research interests include embedded systems and system-on-chip design, security, development support and technology, parallel and energy efficient architectures, computer vision, data analytics, and the application of soft computing and image processing techniques for real world problems. He is a member of the VDE and a Fellow of the IET.