Data augmentation for fairness-aware machine learning

Preventing algorithmic bias in law enforcement systems

Ioannis Pastaltzidis
Information Technologies Institute, Centre for Research and Technology Hellas
gpastal@iti.gr

Nikolaos Dimitriou
Information Technologies Institute, Centre for Research and Technology Hellas
nikdim@iti.gr

Katherine Quezada-Tavárez
Centre for IT and IP Law, KU Leuven
katherine.quezada@kuleuven.be

Stergios Aidinlis
Trilateral Research
stergaidin@gmail.com

Thomas Marquenie
Centre for IT and IP Law, KU Leuven
thomas.marquenie@kuleuven.be

Agata Gurzawska
Trilateral Research
agata.gurzawska@trilateralresearch.com

Dimitrios Tzovaras
Information Technologies Institute, Centre for Research and Technology Hellas
dimitrios.tzovaras@iti.gr

ABSTRACT

Researchers and practitioners in the fairness community have highlighted the ethical and legal challenges of using biased datasets in data-driven systems, with algorithmic bias being a major concern. Despite the rapidly growing body of literature on fairness in algorithmic decision-making, there remains a paucity of fairness scholarship on machine learning algorithms for the real-time detection of crime. This contribution presents an approach for fairness-aware machine learning to mitigate the algorithmic bias / discrimination issues posed by the reliance on biased data when building law enforcement technology. Our analysis is based on RWF-2000, which has served as the basis for violent activity recognition tasks in data-driven law enforcement projects. We reveal issues of overrepresentation of minority subjects in violence situations that limit the external validity of the dataset for real-time crime detection systems and propose data augmentation techniques to rebalance the dataset. The experiments on real world data show the potential to create more balanced datasets by synthetically generated samples, thus mitigating bias and discrimination concerns in law enforcement applications.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Social and professional topics** → Computing / technology policy; • **Surveillance** → Governmental surveillance;

KEYWORDS

computer vision, fairness, algorithmic bias, AI ethics, violence detection, law enforcement technology

ACM Reference Format:
Ioannis Pastaltzidis, Nikolaos Dimitriou, Katherine Quezada-Tavárez, Stergios Aidinlis, Thomas Marquenie, Agata Gurzawska, and Dimitrios Tzovaras. 2022. Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3531146.3534644

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3534644

1 INTRODUCTION

In the last years, Artificial Intelligence (AI) and Machine Learning (ML) systems have become increasingly present in every aspect of life, including the security sector. A promising application in this context is the use of computer vision for the real-time detection and avoidance of crime [64, 65]. However, concerns have been raised about the risks associated with these technologies, with algorithmic bias [20] being one of the most salient ethical, legal and societal challenges for data driven systems. Computer vision algorithms learn to perform a task by capturing relevant characteristics from training data. When trained for seemingly unrelated tasks like activity recognition, models have been found to develop flawed correlations regarding race, gender and age [15, 21]. One of the main reasons for those flawed correlations is datasets representing historic or systemic bias. For instance, violence datasets (which are of high importance for the security sector as they can be used to train computer vision models to detect and deter crime) [63] may lack sufficient representation of all groups, thus likely to lead to misclassification of individuals.

Errors in the outputs of computer vision technology used to perform tasks such as violent activity detection or recognition of suspicious behavior can have serious consequences, e.g., someone being wrongfully arrested based on erroneous but confident misidentification of certain activity as violent and potentially of a

criminal nature. If left unaddressed, dataset bias can result in AI and ML-driven systems perpetuating discrimination and placing certain individuals at a systematic disadvantage. In fact, such pathogenies may not only be replicated by AI systems, but even aggravated by the use, impact and pervasiveness of the technology in society [56]. Given inherent challenges related to ML algorithms and the well-documented concerns of systemic bias in this context [13, 25], it is reasonable to assume that the available training datasets are not balanced in terms of certain attributes, mainly the race, gender and age of the individuals whose data is recorded therein. This is precisely what we encountered in our efforts to develop ML-powered augmented reality technology to improve situational awareness in real-time law enforcement decision-making [12]. In this context, an inherent component to building trustworthy AI and ML systems is the identification and mitigation of unwanted bias to avoid the negative implications of these solutions, bearing in mind that, beyond an ethical requisite, accounting for fairness in AI and ML is a legal requirement stemming from non-discrimination as a core tenet under EU law [38].

The recent interest in fairness-aware ML has resulted in multiple proposals for mitigation strategies to prevent models from learning or applying biases (e.g., [31, 67, 69, 72]; see also [50]). However, only a few works deal with fairness in ML for law enforcement applications [26], with a focus on predictive technology as opposed to crime detection and recognition systems. In this paper, we take a step in this direction by making three contributions. First, we address fairness issues in computer vision for crime detection technology, as opposed to the concerns regarding risk assessment tools used in criminal justice (such as predictive policing and recidivism algorithms) widely considered in the literature. Second, our contribution exposes bias issues in datasets depicting violence situations and suspicious behavior, which serve as raw material for law enforcement technology. Third, we propose a novel strategy to mitigate bias in violence datasets based on data augmentation, a technique to increase the amount of data in the training set by creating (realistic) modified data from the existing corpus [52]. We show how dataset augmentation can be used to introduce controlled biases, which should then eliminate or at least alleviate unfair bias concerns. In that way, the proposed strategy helps address algorithmic bias issues in data-driven policing, thus improving compliance with ethical frameworks and legal norms in the security sector. To this end, we engaged in an interdisciplinary exercise, incorporating the perspectives social scientists (privacy, ethics and legal experts) in a joint effort to implement fairness by design in law enforcement technology.

The rest of the paper is organized as follows. Section 2 analyzes previous work on algorithmic bias in criminal justice and fairness-aware ML. Next, we provide an overview of the fairness standards applicable in the EU in Section 3. Section 4 presents the proposed bias mitigation strategy, describing the dataset used for the experiments and explaining the methodology and limitations of this study. Section 5 details our experimental results, revealing the race and gender bias issues in the dataset, while Section 6 demonstrates the potential of the proposed approach to create more balanced datasets and address bias concerns in law enforcement technology. Section 7 concludes.

2 RELATED WORK

This work is related to a larger body of work on AI in criminal justice and fairness in ML.

Bias, discrimination and fairness in AI for criminal justice. There has been a surge in scholarly and popular interest in algorithmic fairness in recent years. Notably, a substantial body of fairness-related work has emerged from interdisciplinary research communities, with the Fairness, Accountability, and Transparency in ML network (FAccT, formerly FAT) [7] being the most well-known convening venue for those purposes. The closer scrutiny of algorithmic decision-making in recent decades has exposed the tendency of ML systems to embed bias and discrimination into lines of computer code. An example of this is the pioneering study by the Gender Shades project [15], which found significant misclassifications of individuals in ML algorithms based on protected attributes like race and gender. Other examples related to computer vision systems include the study published in 2019 [21], showing how object recognition systems trained on publicly available datasets performed poorly on analysis regarding low-income communities. A more recent example concerns an image labelling service, which classified a thermometer as a "tool" when in the hands of a light-skinned person, while the same object was perceived as a "gun" in a dark-skinned hand [68].

AI, big data and algorithmic systems are proliferating in their use within criminal justice practices given the opportunities for the prevention, detection, investigation and prosecution of crime they create. While these applications hold many promises, they also hold many perils, as shown in the rich literature on the human rights implications AI-powered systems used in criminal justice (e.g., [33, 36, 37, 44, 51]). An example of this is the seminal article by the non-profit media organization ProPublica in 2016 concerning COMPAS [11], an algorithmic system used in bail decisions in the US, where the authors concluded that the technology was biased against black defendants. Another example is the study conducted in the UK, where legal researchers scrutinized the introduction of the Durham Constabulary's Harm Assessment Risk Tool (HART) [55], and cautioned against the risk of unfair outcomes leading to discrimination. In a more recent analysis, Akpinar et al [8] demonstrate how predictive models exclusively trained on victim crime reporting data can yield spatially biased results due to geographic heterogeneity in crime reporting rates.

These applications raise major concerns about the potential aggravation existing problems, such as bias and discrimination, which may result in grave human rights impacts particularly in high stakes decision-making situations like law enforcement. For instance, concerns have been raised about the over-representation of marginalized groups in datasets depicting violent behavior or potentially criminal acts, resulting in the exacerbation of the marginalization of vulnerable groups and engendering higher levels of scrutiny and surveillance into their lives [30]. Another study exposed concerns about both the over- and under-representation of ethnic minorities in ML systems, giving rise to simultaneous implications: the algorithmic tools work poorly on minorities due to the under-representation of these populations in the training datasets, while these same populations are subject to higher levels of surveillance

given historic over-policing practices, turning out to be disproportionately over-represented in training data about crime [14].

The work presented here does not neatly fit into any of the existing publications as it focuses on computer vision systems used for activity classification in law enforcement rather than predictive or risk assessment models, which are widely covered in existing literature.

Fairness-aware machine learning. On the risk mitigation side, prior work has focused on fairness metrics as a way to quantify undesirable bias. This raises questions regarding how social goals are abstracted and employed in prediction tasks [53]. Such metrics are based on the implicit premise that it is possible to establish and operationalize a mathematical concept of fairness to build a system that is devoid of bias [22, 70]. However, Wachter, Mittelstadt, and Russell [66] argue that fairness cannot be automated due to the gap between legal, technical, and organizational notions of algorithmic fairness. In this context, it is important to ensure that technical mitigation measures to address algorithmic bias issues are accompanied by non-technical considerations.

Various discrimination discovery methods have been developed to mitigate fairness concerns. Those methods are sub-divided into *pre-processing* approaches, consisting of manipulation of data or features to allow for a fairer representation of minorities, *in-processing* approaches, entailing the reformulation of the classification problem by considering the discrimination behavior in the optimization function, and *post-processing* approaches, involving the correction of the resulting model [9]. Our work falls under the first category since the goal is precisely to transform "raw" data into a usable form, is one of the steps of the AI process that could result in discriminatory practices, which is the goal pursued with pre-processing methods.

Some authors have developed approaches to transform the training and testing datasets to prevent models from utilizing or learning biases. Research stemming from the early days of bias mitigation attempted to address this concern through techniques for rather simple linear models [48, 70]. More recently, that research has evolved into endeavors concerning more sophisticated models. For instance, a technique presented by Calmon et al [17] consists of reducing the output's reliance on known discriminatory variables (such as race and gender), while at the same time making sure that the resulting outcomes of the system do not significantly differ from the original dataset. Zhao et al [73] introduce an interference update scheme to match a target distribution that can remove bias. Ryu, Adam and Mitchell [58] propose InclusiveFaceNet as a technique to achieve better attribute detection across gender and race subgroups. Dwork et al [23], in contrast, present a scheme based on a decoupling technique that enable to learn different classifiers for different groups. Another relevant approach to bias mitigation is adversarial training [10, 35, 72]. For example, Wang et al [67] use data augmentation to introduce controlled biases in the dataset, with the aim of creating a benchmark for studying bias mitigation.

Most related to our work are earlier studies involving data augmentation techniques as a bias mitigation strategy. To the best of our knowledge, only Iosifidis and Ntoutsi [43] have applied data augmentation techniques, demonstrating its potential to reduce classification error for discriminated groups. Their approach, however, was confined to supervised learning systems. We continue this line of work for bias mitigation methods through data augmentation techniques combined with an active learning architecture, introducing a benchmark for bias evaluation and mitigation of ML applications, specifically addressing the concerns raised by law enforcement technology.

3 ETHICAL AND LEGAL BACKGROUND REGARDING ALGORITHMIC FAIRNESS IN THE EU

Many countries and regions are engaged in a global competition to seize the opportunities brought by AI, and Europe is not the exception. In this context, the EU is investing heavily in AI. However, as is tradition in this region, the efforts are channeled to fostering AI uptake while ensuring safeguards to fundamental rights of people and businesses. In view of this, the EU strives to secure adequate regulation to foster AI that is "trustworthy", that is legal, ethical and robust. For technologists, that means building AI and ML that accounts for fairness, amongst other fundamental and human rights and ethical values.

Fairness has traditionally been conceived as a fundamental ethical principle and a legal foundation of constitutional significance for Member States and the Union. Fairness and the freedom from non-discrimination are fundamental ethical and legal values, enshrined both in legal documents (European constitutions and human rights instruments) and in virtually all the major ethics of AI frameworks [40]. There are different definitions of fairness and non-discrimination [49], mainly converging on the normative requirement that people are provided with equal rights and opportunities, without unjustifiable advantages or disadvantages for certain groups or individuals. Underlying notions of human dignity and respect for human agency underpin the ideals of fairness: human beings are born equal, and their life has an intrinsic worth and value, and they shall be allowed to pursue their freedom and autonomy by being offered the same fundamental rights and opportunities to basic goods like education, health and access to justice.

While this is not indisputable, most legal and ethical frameworks associate fairness and equality with equal resources or opportunities [24], rather than equality of outcomes, e.g., a right to have equal income or status with other people. Discrimination essentially refers to conduct that infringes upon the ethical values of fairness and equality without an ethically defensible justification [34]. More specifically, discrimination involves unequal treatment of individuals or groups due to some of their personal characteristics such as gender, sexual orientation, religion, race, ethnic or national origin, disability or health [47] (the so-called "protected characteristics"). The notion of fairness in a non-discrimination context should be distinguished from fairness under data protection law, as further explained later in this section.

When approaching fairness from an ethical standpoint, recent years have seen the emergence of numerous frameworks of ethical principles, as well as standardization activity on the ethics of AI. Scholars and international organizations have attempted to systematize activity in this space and highlight the most influential frameworks and principles [42, 57]. The EU bodies have been very active in this space, with the most prolific work stemming from the "High Level Expert Group on AI" (HLEG) appointed by the

European Commission. In 2019, the HLEG produced a landmark document for AI ethics, the "Ethics Guidelines for Trustworthy AI" [40]. These Guidelines form part of a broader vision to embrace a human-centric approach to AI, intended to make Europe a global leading innovator in ethical, secure and cutting-edge AI. Accordingly, the guidance document strives to facilitate and enable "Trustworthy AI made in Europe" that will enhance the wellbeing of European citizens [40].

The HLEG Guidelines outline seven fundamental principles for the development of AI systems, among which 'diversity, non-discrimination and fairness' is central. To assist in the implementation of these principles by technology developers and businesses, the HLEG have developed the "Assessment List for Trustworthy AI" (ALTAI), providing more concrete technical design strategies for their implementation in AI systems. The ALTAI is available both as a document and as a prototype of a web-based tool [27]. The HLEG Guidelines highlight the need to avoid unfair bias, to foster accessibility and universal design, meaning that AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, or abilities, and highlight the importance of stakeholder participation throughout the entire process of implementing AI technology.

Standards organizations have also been active in this space, with the Institute of Electrical and Electronics Engineers (IEEE) being a leading actor in this context. Their "Ethically Aligned Design: Prioritizing Human Wellbeing with Autonomous and Intelligent Systems" document, created by "more than 700 global experts" [41], aspires to establish frameworks "to guide and inform dialogue and debate around the non-technical implications' of AI technologies, including considerations beyond individual moral rights such as social fairness, environmental sustainability and self-determination.

Non-discrimination is mandated by EU law too. Of the many legal regimes applicable in an AI context, non-discrimination and data protection law are the two most relevant sources of rules to safeguard against algorithmic discrimination [74]. Discrimination is prohibited in various EU legal instruments and constitutions, including the Charter of Fundamental Rights of the EU (Charter) [1]. Article 21 of the Charter states that "[a]ny discrimination based on any ground such as sex, race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited". Various EU secondary instruments have also been developed, specifically devoted to non-discrimination. The four main non-discrimination directives are: i) the Racial Equality Directive [2]; ii) the Gender Equality Directive [5]; iii) the Gender Access Directive [4]; and iv) the Employment Directive [3]. These instruments establish a minimum standard that needs to be transposed into national law by the Member States.

Two general types of discrimination are addressed under European non-discrimination law. The first is direct discrimination, which relates to adverse treatment based on a protected characteristic on the individual (e.g., gender, race, or sexual orientation). The second type is indirect discrimination, which refers to a situation where "apparently neutral provision, criterion or practice" [2–5] disproportionately disadvantages a protected group in comparison with other individuals. Of those two, it appears that indirect discrimination is the most relevant one in the context of AI and ML applications [38, 74].

Data protection law is also a legal tool that seeks to protect fairness and fundamental rights, including the right to non-discrimination (see General Data Protection Regulation (GDPR), Article 1(2) and Recital 71, 75, and 85). In particular, fairness is one of the principles for the processing of personal data (GDPR, Article 5(1)(a)), and thus sits at the core of EU data protection law. However, fairness under data protection law should not be confused with fairness in the non-discrimination law context. Under non-discrimination law, fairness measures or initiatives are aimed at enforcing equality. For instance, algorithmic fairness research "seeks to operationalize equality at a mathematical level" [38]. In a data protection context, fairness is to be understood as processing personal data in ways that individuals would reasonably expect and not doing so in a manner that could result in unjustified adverse effects for persons. Yet, the conditions set forth under data protection law can mitigate possible risks to the rights to fairness and non-discrimination [38]. For example, data protection imposes the obligation of transparency of data processing activities, which entails the need to provide information to data subjects about AI-assisted decision-making involving personal data (GDPR, Articles 5(1)(a), 13 and 14). There is also the obligation to conduct a data protection impact assessment under certain circumstances, for instance when using new technologies (GDPR, Article 35(3)(a)).

It is also worth mentioning the forthcoming EU regulation aiming to establish a legal framework for trustworthy and fair AI. Specifically, in April 2021, the EU published its Proposal for a Regulation laying down harmonized rules on AI [6], becoming the first political entity to officially put forward a formal initiative on AI regulation [28]. This legal instrument will be applicable to providers and users of AI systems, following a risk-based approach, where different rules apply to different risk levels of AI: the higher the risk posed by the relevant system, the stricter the rules that will apply. Amongst the obligations related to high-risk systems, which is likely to be the case for most law enforcement technology, is that of ensuring high quality of the datasets feeding the system to minimize risks and discriminatory outcomes (AI Act, Recital 44 and Article 10).

According to a study published in 2021, it is currently not possible to automate fairness or non-discrimination, particularly "because the law does not provide a static or homogenous framework suited to testing for discrimination in AI systems" [66]. Yet, it is still possible to take certain measures towards ensuring that the fairness policies are satisfied. For instance, considering the importance of the choice of the data used [29], it is important to ensure that the datasets used to train, validate and test the AI system are relevant, representative, free of errors and complete. In this respect, when acquiring, labelling, cleaning and preparing datasets, it is important to assess the quality of the data, which starts with a basic understanding of it (e.g., where the data come from, what the dataset covers, which skews and correlations the data contain). The issues to consider when assessing the quality of the data include questions about their completeness, accuracy, consistency, timeliness, duplication, validity, availability and provenance [16]. It is precisely in this context where the data augmentation solution discussed herein

could play a key role in the generation of accurate representations of reality by eliminating or reducing biases reflected in ML training data.

4 COMPONENTS OF THE PROPOSED STRATEGY

Proceeding from the legal and ethical conceptions of fairness and non-discrimination, we co-designed a data augmentation strategy for mitigating algorithmic bias risk in the context of an EU-funded research and innovation action that develops ML-powered smart devices for real-time crime detection [12]. Data augmentation consists of the process of generating data through the use of information from the training dataset. With this process, it is possible to synthetically oversample under sampled entities to re-balance the dataset. Through the data augmentation methodology, we synthetically re-balance training datasets to mitigate bias and discrimination risks without compromising the integrity and reliability of data. A typical example would be to augment a training dataset so that race and gender percentages are more balanced, as done in this study. This section explains the components of this approach.

4.1 Dataset

The dataset that we used for this experiment is RWF-2000, which contains two thousand video clips lasting five seconds each, half of them depicting violent acts and half of them normal activities. We chose to evaluate this dataset given its importance as raw material for law enforcement technology since it is "largest surveillance video dataset used for violence detection in realistic scenes" [19]. The dataset is split in 80% for training and 20% for evaluation. Preliminary analysis of the dataset revealed overrepresentation of dark-skinned males in violence situations, underrepresentation of light-skinned males, regarding the general population of a country (specifically the United States) and underrepresentation of females in general. Before explaining the composition of the dataset, a caveat is in order. There is underrepresentation of darker males in the videos with regards to the general distribution of the "race" category. However, in the dataset the ratio white/black is 2:1.

In our evaluation of the RWF-2000 dataset, we found that 81% of the people in violent acts in the training set were men and 19% women, while the non-violent training samples had 65% men and 35% women. For the evaluation set the distribution for men was 81% and for women 19% on violent acts and for the non-violent acts was 69% to 31% respectively. Race distribution in violent training samples was 37% white, 23% black, 20% Asian while other groups constituted the rest, in non-violent training samples the distribution was 36% white, 18% black and 20% Asian. In violent evaluation samples the distribution was 36% white, 22% black and 25% Asian, and in nonviolent samples 44% white, 18% black and 15% Asian. We applied data augmentation techniques to achieve better representation of violent acts and normal activities on the basis of race.

4.2 Methodology

In this study, we apply data augmentation techniques to artificially create instances based on a given attribute (focusing on race in this case). In that way we force balance by populating the underrepresented group for a specific attribute to deal with group's class imbalance. Specifically, our proposed approach is to replace a person in a violent or non-violent video sequence with another person, imitating the individual's movement throughout the sequence. In order to achieve this, we use Mask-RCNN [39] for instance segmentation and calculate the entity's keypoints with HRNet-w48 [60]. Instance segmentation is an object detection technique that generates a segmentation map for each detected instance of an object in an image. Pose estimation refers to the detection of keypoints (neck, eyes, wrists, etc.) in an instance of a person. Mask-RCNN is a method that predicts bounding boxes along with the masks, consequently these are the bounding boxes that we use. The joints' locations are estimated by re-adjusting the bounding boxes for the person to match the pose estimation model's aspect ratio, crop the image in accordance with the re-adjusted bounding boxes and interpolate it to the model's input resolution. After processing the 17 heatmaps produced from the pose estimation model (corresponding to 17 joints), the keypoints are projected back to the original image coordinates Examples of instance segmentation and pose estimation on RWF-2000 are provided in Figure 1.

To ensure that we are replacing the same person in each video, we track the persons throughout the video sequence. We also extract the keypoints of the person that will replace the target individual. We then scale their bounding boxes to make them equal in size and transform their keypoints and masks to the new pixel coordinates. This is done with affine transformation matrices for scaling and translation. In this case, the affine transformation matrices will look like the 3x3 matrix in the equation 1, where x' and y' are the new pixel coordinates, x, y are the original pixels, s_x, s_y are the scaling factors for x and y and t_x, t_y are the translation distances. We provide the bounding box of the person we want to replace and the bounding box of the one who is replacing and solve the system (4.2), where $[(x'_1, y'_1), (x'_2, y'_2)]$ are the bounding box coordinates for the person we want to replace and $[(x_1, y_1), (x_2, y_2)]$ the bounding box coordinates of the person we are replacing with, to obtain the scaling factors and translation distance. Examples of scaling and transformation are provided in Figure 2. From left to right, the first two images depict the person we are going to warp and his keypoint location, image 3 is person we are replacing, and the last images provide the scaled person and his transformed keypoints.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

$$\begin{aligned} x'_1 &= s_x * x_1 + t_x \\ x'_2 &= s_x * x_2 + t_x \\ y'_1 &= s_y * y_1 + t_y \\ y_2 &= s_y * y_2 + t_y \end{aligned} \tag{2}$$

We then calculate the displacement vectors for each respective body part by subtracting the keypoints of the person we track through the video sequence from the transformed keypoints of the replacement person, which are calculated using the 3x3 Affine Transformation matrix of equation 1 after having already calculated the scaling and translation distances from the system in equation 2. We use Radial Basis Function (4.3) for interpolation. We have 17

Figure 1: From left to right, the first two images are examples of instance segmentation, the next two examples of pose estimation.

Figure 2: Scaling and Transformation example.

data points in total corresponding to the joints' location and the displacement vector as their values for x and y, thus two approximation functions are initialized from equation 3, 1 for the x pixels and 1 for the y pixels. We generate a mesh grid, consisting of the image pixel coordinates, which will be used as input to these two approximation functions for estimating the displacement vector for each x and y pixel. We interpolate and generate the displacement field. Since we do not want to alter the background, we use the mask to nullify the field outside of the mask's boundaries. We then warp the replacement entity in a way that the individual resembles the pose and motion of the person originally in the video. Thus, we can generate synthetic video sequences that can be more balanced than the original ones.

$$f(x) = \sum_{i=1}^{N} w_i \varphi(x - x_i) \tag{3}$$

Our data augmentation pipeline is depicted in Figure 3. Figure 4 shows an example of our technique, through various images representing the different steps. From left to right, starting on the top row, Image 1 is the original frame, and Image 2 represents the original keypoints of the replacement person. Image 3 represents the calculated keypoints for the target entity (i.e., the person tracked throughout the video frames and subject to replacement), and Image 4 shows the transformed keypoints location. Images 5 and 6 show the corresponding displacement fields for X and Y pixels, respectively. Lastly, Image 7 is the resulting altered frame.

Having rebalanced the datasets, we perform some experiments to evaluate the effects of the data augmentation strategy on the model's performance in terms of fairness.

4.3 Limitations

One issue that we have observed is that most of the videos in the dataset suffer from poor quality and lighting, limiting the available video clips that we can use. Furthermore, in violent events where people are engaged in close fights and are entangled with each other, Mask-RCNN has difficulties distinguishing them and returns arbitrary shaped masks or one mask for the two individuals. Hence, the replacement in some frames might be sub-optimal. We have also observed that warping sometimes fails to align limbs' position, particularly is cases where the hands or legs must be moved/raised many pixels across the image. Figure 5 shows some failure cases. From left to right, the first column of images depicts the person we are tracking and replacing, the second column shows the replacement person and the third column the resulting altered frames. In the first image in the third column, Mask-RCNN did not detect the tracked person, so no replacement was performed. In the second image of the same column the hand of the person we are tracking was not detected and the warping was in general not ideal. Lastly, in the third image in the third row, the original person is still visible.

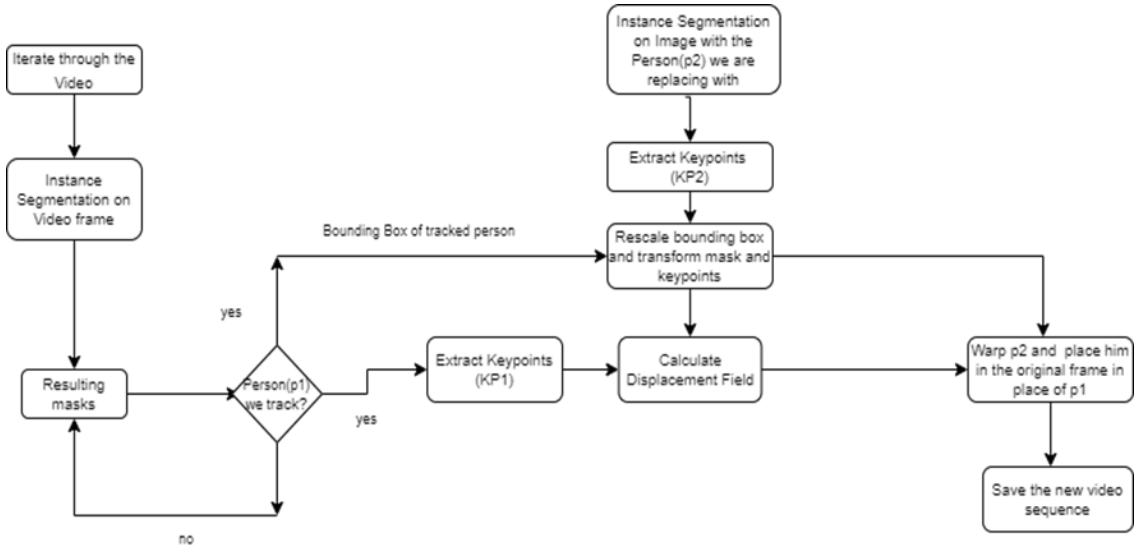

Figure 3: Data Augmentation Pipeline

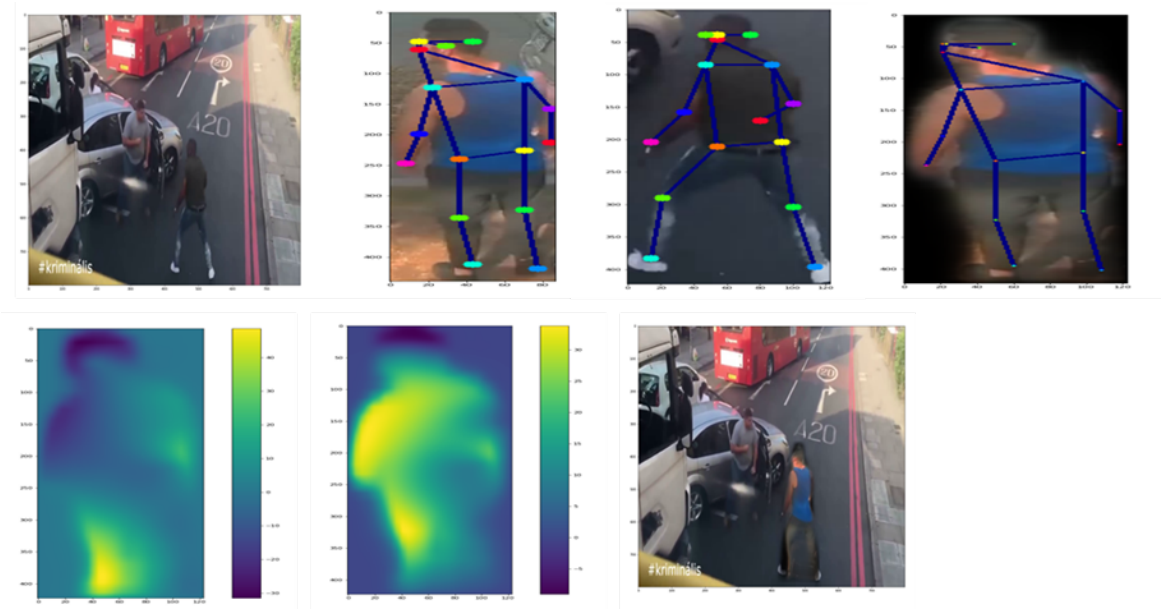

Figure 4: An example of person replacement in a video frame. In the first row, from left to right, Image 1 is the original frame, Image 2 depicts the person who is replacing, Image 3 the person who is getting replaced, Image 4 is the person who is replacing scaled. In the bottom row, from left to right Images 5 and 6 are the displacement fields for x and y and Image 7 is the resulting altered frame.

5 EVALUATING THE MODEL

The models that we used for our experiment is RTFM [61], which has reported state-of-the-art metrics across different datasets, and Sultani et al [59] method. Both these methods use a pre-trained feature extractor and a classifier. Sultani et al use C3D [62], while RTFM uses I3D [18]. C3D is pre-trained on the Sports-1M [45] dataset, while I3D on Kinetics-400 [46] or Charades [70] (the one we used was pre-trained on Charades). The results presented in this and the next section are on per 24 frame segments for RTFM and on per 16 frame segments for Sultani et al [59]. We use Pytorch, torchvideo, numpy, OpenCV and sci-kit learn for implementation and experimentation. Initially, we ran some experiments with RWF-2000, which led to the conclusion that the model's performance is

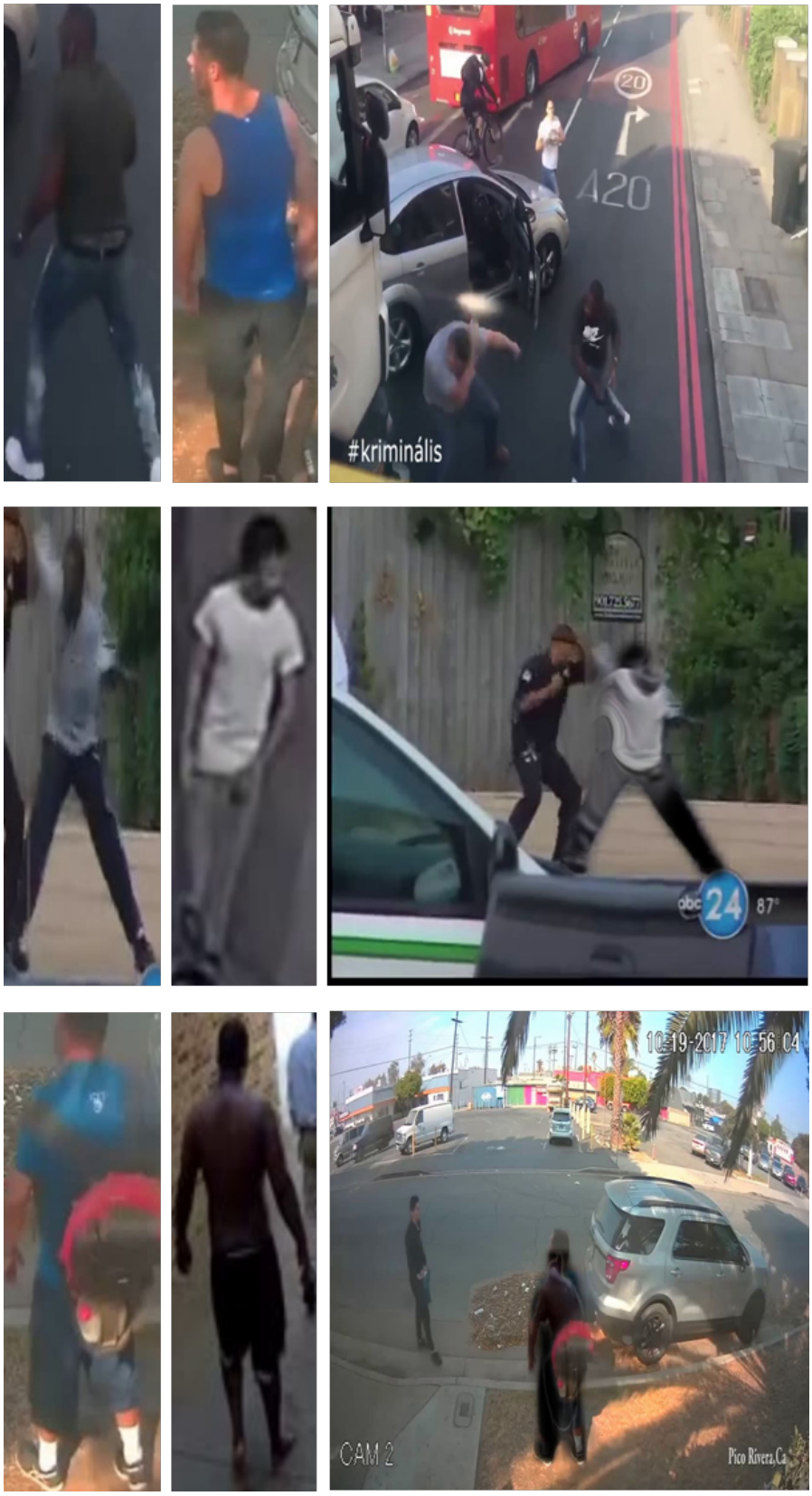

Figure 5: Failure cases.

Table 1: RTFM Results on RWF-2000 gender

Models	Test Set	Mean AUC	Mean AUC std	Mean Accuracy	Mean Accuracy std
Balanced models	Balanced	*0.878*	0.003	*0.772*	0.026
Balanced models	Imbalanced	0.761	0.005	0.671	0.027
Imbalanced models	Balanced	0.847	0.009	0.766	0.020
Imbalanced models	Imbalanced	*0.847*	0.006	*0.766*	0.029

Table 2: Sultani et al method on RWF-2000 gender

Models	Test Set	Mean AUC	Mean AUC std	Mean Accuracy	Mean Accuracy std
Balanced models	Balanced	*0.814*	0.016	*0.727*	0.033
Balanced models	Imbalanced	0.814	0.023	0.721	0.030
Imbalanced models	Balanced	0.729	0.020	0.647	0.020
Imbalanced models	Imbalanced	*0.861*	0.015	*0.765*	0.029

affected by gender and race balance issues, as shown in Table 1 and Table 3.

Balanced and imbalanced model refers to models that were trained with balanced or imbalanced datasets based on gender or race. We resampled the training samples and the male to female ratio was around 1:1 in the balanced dataset, while in the imbalanced dataset the ratio was 4:1, which was the same as in the original RWF-2000 corpus. The same ratios were kept for the balanced and imbalanced test datasets. Supposing we have positive (P) and negative (N) samples a binary classifier can have four outcomes, True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN). For the purposes of this analysis, accuracy is to be understood as follows:

$$Accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FP + TN + FN}$$

Area under the curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one [32].

The Standard Deviation (std) is defined as:

$$std = \sqrt{\frac{1}{N} * \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

, where N is the sample size, $\bar{x}$ is the mean value of the samples and x_i is the value of sample i.

In this experiment, we deployed ten-fold cross validation schemes. Since the dataset does not have annotations for the classes contained therein, one author annotated the dataset with race and gender labels of the people present in the video clips, using female and male labels to indicate subjects perceived as women or men respectively, and white and black labels to indicate subjects perceived as lighter skinned and darker skinned respectively. We had also tested a different model from Sultani et al and the results are presented in Table 2.

We also experimented with race imbalance in the RWF-2000 dataset. The original dataset had a white to black ratio of 1:6 in violent videos for training and 2:0 ratio on non-violent videos. We created 2 sub-datasets, one with 2:0 ratio and one with 1:0 ratio, which we call "imbalanced" and "balanced" respectively, and ran 2 5-fold Cross-Validation schemes. The results are presented in Table 3. Table 4 provides statistics for the gender and race distribution for the datasets used for training.

Table 1 shows that the RTFM models that were trained on imbalanced training samples had similar performance when they were tested on balanced and imbalanced data, as their mean AUC score stood at 84.7% and mean accuracy at 76.6%, while those which were trained on balanced training samples performed better on the balanced test dataset (87.8% AUC and 77.2% accuracy) but had a significant drop in their metrics when tested on the imbalanced test dataset (76.1% and 66.1% for the AUC and accuracy, respectively). Table 2 shows that the models that were trained on balanced dataset performed similarly when tested on balanced and imbalanced samples as their AUC score was the same and the accuracy was slightly lower on the imbalanced test set, while the models that were trained on imbalanced data performed well on imbalanced test samples and poorly on balanced test samples, as their AUC and accuracy reduced significantly. Table 3 shows that a model that is trained on more balanced dataset is able to achieve higher accuracy scores at certain thresholds and generalize better. The models that were trained on balanced training data had a mean accuracy of 67.7% for the balanced test set and 73.6% for the imbalanced test set, while those which were trained on imbalanced videos had 66.5% and 72.2% for the balanced and imbalanced test set respectively.

When evaluating these results, it is evident that they do not always meet the expectations. We would generally expect that the models that were trained on balanced datasets would perform better than the ones trained on imbalanced datasets. In the results presented in Table 1, the balanced models perform better on the balanced test set, as one would expect, but the imbalanced models are the ones that can generalize better since they have similar performance on the balanced and imbalanced test sets. In the second experiment (Table 2), the balanced models are the ones that generalize better as they have the same performance for balanced and imbalanced test sets, while the imbalanced ones perform well on the imbalanced test set and poorly on the balanced one. These

Table 3: RTFM method on RWF-2000 race

Models	Test Set	Mean AUC	Mean AUC std	Mean Accuracy	Mean Accuracy std
Balanced models	Balanced	0.793	0.006	*0.677*	0.038
Balanced models	Imbalanced	0.854	0.005	*0.736*	0.025
Imbalanced models	Balanced	*0.805*	0.005	0.665	0.050
Imbalanced models	Imbalanced	*0.866*	0.007	0.722	0.047

Table 4: Gender and race distribution on Training datasets

Dataset	Males	Females	Light-Skinned	Dark-Skinned	Other
Balanced (gender)	50.7%	49.3%	37.3%	22.9%	39.8%
Imbalanced (gender)	80.5%	19.5%	25.2%	16.9%	57.9%
Balanced (race)	74.2%	25.8%	79.4%	17%	3.6%
Imbalanced (race)	74.1%	25.9%	63.9%	31%	5.1%

inconsistencies could be attributed to the fact that during Cross Validation each fold might have a distribution that is not representative of the datasets that are used for testing.

Considering these results, it can be concluded that gender and race balance affects the models' performance. We have shown that a racially balanced training set leads to higher accuracy and in turn better generalization than an imbalanced training set and a gender balanced training set can either increase or decrease the model's performance depending on the method.

6 FIXING REAL-TIME CRIME DETECTION SYSTEMS

The data augmentation approach stems from the results presented in Section 5. More specifically, we focused on the issue of race imbalance identified in the data. The approach is rooted in the idea that, if we generate more balanced training samples, the model will be able to better generalize the action instead of the person doing it. We used a subset of RWF-2000 which had dark-skinned individuals in almost every violent video and few in non-violent videos. We replaced a dark-skinned individual in the violent scenes mainly with light-skinned persons and light-skinned individuals mainly with dark-skinned persons in non-violent scenes. We also tried to have individuals appearing in fight scenes to be present in non-violent scenes as well, and vice versa.

By applying data augmentation, we created twenty new videos that looked promising and good enough, ten violent and ten non-violent, thus creating a new dataset with more balanced skin type representations with our technique. We then deployed five-fold cross validation schemes to determine whether this type of data augmentation had any effect on the model's accuracy. In the first cross validation we used the twenty original videos for training while in the second we had the original videos and the synthetic ones, forty in total, as training examples. Tests have shown that while the AUC was 1% lower to 1.5% higher for the models that were trained on both the original and synthetic videos in different test datasets compared to those which were trained on only the original videos, the models that were trained on both the original and the synthetic samples had higher accuracy at practical thresholds. For example, with a threshold of 0.4 the accuracy was 2 to 4 percent higher. Comparative results are presented in Table 5 and Table 6.

Table 5 contains three test dataset: one with black majority, one with white majority and lastly one with black to white ratio 1:1. We trained each model for 90 epochs and tested them on the three test datasets. Comparing the results, it becomes evident that the models that were trained on both the original and the synthetic videos have higher accuracy and the difference in accuracy between the datasets decreases. For example, the models that were trained only on the original videos had a mean accuracy of 0.589 on the white majority dataset and only 0.562 on the black majority dataset, a difference in accuracy of 2.7% while the models that were trained on both the original and the synthetic videos had 0.618 and 0.600 respectively on these two datasets, a difference of 1.8%.

Table 6 shows the results for 150 epochs of training. Here again the higher accuracy of the models that were trained on both the original and synthetic videos is apparent. For the black majority dataset, the models that were trained on the original videos had an accuracy of 0.610, while in the white majority the accuracy was 0.652, meaning that there is a difference of 4.2% between the two. The models that were trained on both the original and synthetic video datasets had 0.633 and 0.664 accuracy in black and white majority test set respectively, a difference of 3.1%.

In light of these results, it can be concluded that our data augmentation does not only have the potential to mitigate bias in the model, but it can also increase the model's accuracy in terms of practical accuracy thresholds. Nonetheless, it is important to acknowledge some downsides of this method, particularly the fact that the processing of the frames is done at a rate of 3 fps on a laptop with Nvidia GTX 1660 Ti. Yet, this rate should not affect the model's inference during testing since this data augmentation technique is meant to be used during the model's training phase.

7 CONCLUSION

In this work, we deal with bias towards certain attributes in violence datasets. We propose an approach based on data augmentation techniques to create pseudo instances to have more balanced training samples. We found that when rebalancing the dataset in regard to

Table 5: RTFM on RWF-2000 race bias 90 epochs training

Model's training	Test Set	Mean AUC	Mean AUC std	Mean Accuracy	Mean Accuracy std
Original Videos	Black Majority	*0.698*	0.045	0.562	0.039
Original Videos	White Majority	0.746	0.026	0.589	0.059
Original Videos	Balanced	*0.703*	0.026	0.567	0.044
Original & Synthetic Videos	Black Majority	0.6922	0.013	*0.600*	0.048
Original & Synthetic Videos	White Majority	*0.760*	0.028	*0.618*	0.049
Original & Synthetic Videos	Balanced	0.695	0.017	*0.607*	0.038

Table 6: RTFM on RWF-2000 race bias 150 epochs training

Model's training	Test Set	Mean AUC	Mean AUC std	Mean Accuracy	Mean Accuracy std
Original Videos	Black Majority	*0.703*	0.027	0.610	0.054
Original Videos	White Majority	0.765	0.024	0.652	0.063
Original Videos	Balanced	*0.709*	0.006	0.620	0.052
Original & Synthetic Videos	Black Majority	0.690	0.020	*0.633*	0.042
Original & Synthetic Videos	White Majority	*0.782*	0.030	*0.664*	0.054
Original & Synthetic Videos	Balanced	0.705	0.020	*0.642*	0.028

race and training new models with such data the resulting models tend to perform better (1.2% higher accuracy in a balanced test dataset and 1.4% higher accuracy in imbalanced test dataset) than the models trained with an imbalanced dataset (compared to the original ratio for lighter skinned and darker skinned people). Our experiments have shown that the models which are trained on balanced datasets are able to generalize better, with some exceptions. In any case, dataset balance during training plays a crucial role in the model's performance and its ability to generalize.

The gained insights in this study show that the data augmentation method applied in this research, consisting of the use of displacement fields and warping, is able to increase the model's accuracy while making the model less biased on race. In particular there was an increase in accuracy compared to the model that was not trained on the augmented data across all 3 test datasets. Regarding the AUC score, it was increased in 1 out of the 3 test datasets, the white majority one, and it had marginal losses in the other 2 test datasets. To summarize the results, this method indeed seems to increase the classifier's accuracy regardless of the test dataset, in regards to the AUC score it remained basically the same with the exception of the white majority dataset where an increase was observed. The proposed augmentation technique for bias mitigation has promising results as it appears to increase the model's generalization ability, although it has limitations in replacing a person's movements, detecting a person for the entire duration of the video, detecting two masks for the same person or masks that are not fully covering the person when the video quality and lighting are not good enough. Despite the limitations faced in this research, this work can serve as the basis for more advanced techniques on bias mitigation in AI systems in the law enforcement sector.

Given the growing importance of computer vision technology in high stakes situations, such as law enforcement, further research is needed to further improve the bias issues of the datasets used in this context. Motion transfer with the use of Generative Adversarial Networks (GANs) can be used in such task in order to create large quantities of synthetic data which are going to be balanced based on race, gender, age etc. Such techniques could also be deployed on the edge in order to offer online training for AI models which are going to be unbiased on the site that they are present.

ACKNOWLEDGMENTS

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 883297. European Union Institutions had no role in the design and conduct of the study; access and collection of data; analysis and interpretation of data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. The authors declare no other financial interests.

REFERENCES

[1] Charter of Fundamental Rights of the European Union. 2012. OJ C 326, 26 October 2012.
[2] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, OJ L 180, 19.7.2000, 22–26.
[3] Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation, OJ L 303, 2.12.2000, 16–22.
[4] Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services, OJ L 373, 21.12.2004, 37–43.
[5] Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation, OJ L 204, 26.7.2006, 23–36.
[6] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending certain Union Legislative Acts. Brussels, 21.4.2021, COM/2021/206 final.
[7] ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT). https://facctconference.org/
[8] Nil-Jana Akpinar, Maria De-Arteaga, and Alexandra Chouldechova. 2021. The effect of differential victim crime reporting on predictive policing systems. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA,

838–849. https://doi.org/10.1145/3442188.3445877
[9] Kiana Alikhademi, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves, and Juan E. Gilbert. 2021. A review of predictive policing from the perspective of fairness. Artif Intell Law (April 2021), 1–17. https://doi.org/10.1007/s10506-021-09286-4
[10] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaker. 2018. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. arXiv:1809.02169 [cs.CV] (September 2018).
[11] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. ProPublica (May 2016). https://www.propublica.org/article/machine-bias-riskassessments-in-criminal-sentencing
[12] Konstantinos C. Apostolakis, Nikolaos Dimitriou, George Margetis, Stavroula Ntoa, Dimitrios Tzovaras, and Constantine Stephanidis. 2021. DARLENE – Improving situational awareness of European law enforcement agents through a combination of augmented reality and artificial intelligence solutions [version 1; peer review: 2 approved with reservations]. Open Research Europe 1, 87 (July 2021) https://doi.org/10.12688/openreseurope.13715.1
[13] Alexander Babuta, Marion Oswald, and Christine Rinik. 2018. Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges. Whitehall Report No. 18, Volume 3. Royal United Services Institute for Defence and Security Studies and University of Winchester.
[14] Ruha Benjamin. 2019. Race After Technology: Abolitionist Tools for the New Jim Code. Polity Press.
[15] Joy Buolamwini, and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research), Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, 77–91. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a.html
[16] Andrew Burt, Brenda Leong, Stuart Shirrell, and Xiangnong (George) Wang. 2018. Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models. The Future of Privacy Forum White Paper. https://fpf.org/wp-content/uploads/2018/06/Beyond-Explainability.pdf
[17] Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. (2017). Optimized data pre-processing for discrimination prevention. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS'17).
[18] Joao Carreira, and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 6299–6308.
[19] Ming Cheng, Kunjing Cai, Ming Li. 2019. RWF-2000: An Open Large Scale Video Database for Violence Detection. arXiv:1911.05913v3 [cs.CV] (October 2020)
[20] David Danks and Alex John London. 2017. Algorithmic Bias in Autonomous Systems.. In Proceedings of the 26th International Joint Conference on Artificial Intelligence ((IJCAI'17), 4691–4697. https://doi.org/10.24963/ijcai.2017/654
[21] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone?. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '19). IEEE, Long Beach, CA, 52–59. Retrieved from https://openaccess.thecvf.com/content_CVPRW_2019/papers/cv4gc/de_Vries_Does_Object_Recognition_Work_for_Everyone_CVPRW_2019_paper.pdf
[22] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12). Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255
[23] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency. (Proceedings of Machine Learning Research), Vol. 81. PMLR, 119–133. Retrieved from http://proceedings.mlr.press/v81/dwork18a.html
[24] Ronald Dworkin. 1981. What is equality? Part two: Equality of resources. Philosophy & Public Affairs, Vol. 10, 283–345.
[25] Devin English, Lisa Bowleg, Ana Maria del Río-González, Jeanne M. Tschann, Robert Agans, and David J. Malebranche. 2017. Measuring Black Men's Police-based Discrimination Experiences: Development and Validation of the Police and Law Enforcement (PLE) Scale. Cultur Divers Ethnic Minor Psychol 23, 2 (April 2017), 185–199. https://doi.org/10.1037/cdp0000137
[26] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway Feedback Loops in Predictive Policing. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research), Vol. 81. PMLR, 160–171. Retrieved from https://proceedings.mlr.press/v81/ensign18a.html
[27] European Commission. 2020. Artificial intelligence: The Commission welcomes the opportunities offered by the final Assessment List for Trustworthy AI (ALTAI). European Commission (July 2020). https://ec.europa.eu/digital-single-market/en/news/artificial-intelligence-commission-welcomes-opportunities-offered-final-assessment-list
[28] European Commission. 2021. Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682
[29] European Union Agency for Fundamental Rights. 2019. Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights.
[30] Virginia Eubanks. 2018. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press, New York.
[31] Tobias Fahse, Viktoria Huber, and Benjamin van Giffen. 2021. Managing Bias in Machine Learning Projects. Lecture Notes in Information Systems and Organization (WI '21), Vol 47. Springer, Cham. https://doi.org/10.1007/978-3-030-86797-3_7
[32] Tom Fawcett. 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27, 8 (June 2006), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
[33] Andrew Guthrie Ferguson. 2017. The Rise of Big Data Policing. NYU Press. https://doi.org/10.2307/j.ctt1pwtb27
[34] Sandra Fredman. 2016. Emerging from the Shadows: Substantive Equality and Article 14 of the European Convention on Human Rights. Human Rights Law Review, 16, 2 (June 2016), 273–301. https://doi.org/10.1093/hrlr/ngw001
[35] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, Vol. 37 (ICML'15). JMLR.org, 1180–1189.
[36] Gloria González Fuster. 2020. Artificial Intelligence and Law Enforcement - Impact on Fundamental Rights. Report No. PE 656.295, requested by the LIBE committee. Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament. Retrieved from https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2020)656295
[37] Oskar Josef Gstrein, Anno Bunnik, and Andrej Zwitter. 2019. Ethical, Legal and Social Challenges of Predictive Policing. Católica Law Review (December 2019), 77–98. Retrieved from https://research.rug.nl/en/publications/ethical-legal-and-social-challenges-of-predictive-policing
[38] Philipp Hacker. 2018. Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law. Common Market Law Review 55,4 (August 2018), 1143–1185. Retrieved from https://kluwerlawonline.com/JournalArticle/Common+Market+Law+Review/55.4/COLA2018095
[39] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask-RCNN. arXiv:1703.06870 [cs.CV] (March 2017).
[40] High Level Expert Group on AI. 2019. Ethics Guidelines for trustworthy AI.
[41] Institute of Electrical and Electronics Engineers (IEEE). 2019. Ethically Alligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.
[42] Marcello Ienca, and Effy Vayena. 2020. AI Ethics Guidelines: European and Global Perspectives -Provisional report. Report No. CAHAI(2020)07-fin. Ad Hoc Committee on AI, Council of Europe.
[43] Vasileios Iosifidis, Eirini Ntoutsi. (2018). Dealing with Bias via Data Augmentation in Supervised Learning Scenarios. In Proceedings of the Workshop on Bias in Information, Algorithms, and Systems Sheffield, Vol. 2103. United Kingdom, March 25, 2018.
[44] Fieke Jansen. 2018. Data Driven Policing in the Context of Europe. Data Justice Lab Working Paper. Cardiff University. Retrieved from https://datajusticeproject.net/wp-content/uploads/sites/30/2019/05/Report-Data-Driven-Policing-EU.pdf
[45] Andrej Karpathy and George Toderici and Sanketh Shetty and Thomas Leung and Rahul Sukthankar and Li Fei-Fei (2014). Large-scale Video Classification with Convolutional Neural Networks. (CVPR 2014).
[46] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, Andrew Zisserman (2017). The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs.CV] (May 2017).
[47] Tarunabh Khaitan. 2015. A Theory of Discrimination Law. Oxford University Press.
[48] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. 2012. Undoing the Damage of Dataset Bias. European Conference on Computer Vision (ECCV), 158–171.
[49] Kasper Lippert-Rasmussen. 2013. Born Free and Equal? A Philosophical Inquiry into the Nature of Discrimination. Oxford University Press.
[50] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (July 2022), 1–35. https://doi.org/10.1145/3457607
[51] Albert Meijer, and Martijn Wessels. 2019. Predictive Policing: Review of Benefits and Drawbacks. International Journal of Public Administration 42, 12 (February 2019), 1031–1039. https://doi.org/10.1080/01900692.2019.1575664
[52] Agnieszka Mikołajczyk, and MichałGrochowski. 2018. Data augmentation for improving deep learning in image classification problem. In Proceedings of the International Interdisciplinary PhD Workshop (IIPhDW '18). IEEE, Świnoujście, Poland, 117-122. https://doi.org/10.1109/IIPHDW.2018.8388338
[53] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, Kristian Lum. 2018. Prediction-based decisions and fairness: a catalogue of choices, assumptions, and definitions. arXiv:1811.07867v3 [stat.AP] (April 2020).

[54] Organization for Economic Co-operation and Development (OECD). 2019. OECD AI Principles. Retrieved from https://oecd.ai/en/ai-principles
[55] Marion Oswald, Jamie Grace, Sheena Urwin, and Geoffrey C. Barnes. 2018. Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. ICT Law 27, 2 (April 2018), 223–250. https://doi.org/10.1080/13600834.2018.1458455
[56] Safiya Umoja Noble. 2018. Algorithms of oppression. New York University Press.
[57] Raymond Perrault, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles. 2019. The AI Index 2019 Annual Report. t Stanford University's Human-Centered Artificial Intelligence Institute.
[58] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. 2017. Inclusivefacenet: Improving face attribute detection with race and gender diversity. arXiv:1712.00193v3 [cs.CV] (July 2018).
[59] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world Anomaly Detection in Surveillance Videos. arXiv:1801.04264v3 [cs.CV] (February 2019).
[60] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. arXiv:1902.09212 [cs.CV] (February 2019).
[61] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. 2021. Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning. arXiv:2101.10030 [cs.CV] (January 2021).
[62] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2014. C3D: Generic Features for Video Analysis. arXiv:1412.0767 [cs.CV] (December 2014).
[63] Rajesh Kumar Tripathi, Anand Singh Jalal, and Subhash Chand Agrawal. 2018. Suspicious human activity recognition: a review. Artif. Intell. Rev. 50, 2 (August 2018), 283–339. https://doi.org/10.1007/s10462-017-9545-7
[64] Vikas Tripathi, Ankush Mittal, Durgaprasad Gangodkar, Vishnu Kanth. (2019). Real time security framework for detecting abnormal events at ATM installations. J Real-Time Image Proc 16 (April 2019), 535–545. https://doi.org/10.1007/s11554-016-0573-3
[65] Waseem Ullah, Amin Ullah, Ijaz Ul Haq, Khan Muhammad, Muhammad Sajjad and Sung Wook Baik. 2021. CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. Multimedia Tools and Applications 80, 11 (May 2021), 16979–16995. https://doi.org/10.1007/s11042-020-09406-3
[66] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Computer Law & Security Review 41 (2021), 105567. https://doi.org/10.1016/j.clsr.2021.105567
[67] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '20). 8919–8928. Retrieved from https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_Towards_Fairness_in_Visual_Recognition_Effective_Strategies_for_Bias_Mitigation_CVPR_2020_paper.html
[68] Kyle Wiggers. 2020. Researchers show that computer vision algorithms pretrained on ImageNet exhibit multiple, distressing biases. Venture Beat (November 2020). https://venturebeat.com/2020/11/03/researchers-show-that-computer-vision-algorithms-pretrained-on-imagenet-exhibit-multiple-distressing-biases/
[69] Douglas Yeung, Inez Khan, Nidhi Kalra, and Osonde A. Osoba. 2021. Identifying Systemic Bias in the Acquisition of Machine Learning Decision Aids for Law Enforcement Applications. Report No. PE-A862-1. RAND Corporation.
[70] Yuan Yuan and Xiaodan Liang and Xiaolong Wang and Dit-Yan Yeung and Abhinav Gupta (2017). Temporal Dynamic Graph LSTM for Action-driven Video Object Detection. (ICCV 2017)
[71] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In Proceedings of the 30th International Conference on Machine Learning, Vol. 28, 3. PMLR, 325–333. Retrieved from http://proceedings.mlr.press/v28/zemel13.html
[72] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). Association for Computing Machinery, New York, NY, USA, 335–340. https://doi.org/10.1145/3278721.3278779
[73] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17). Association for Computational Linguistics, Copenhagen, Denmark, 2979–2989. https://doi.org/10.18653/v1/D17-1323
[74] Frederik Zuiderveen Borgesius. 2018. Discrimination, artificial intelligence, and algorithmic decision-making. Council of Europe.