Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-15621-5

Deep labeller: automatic bounding box generation for synthetic violence detection datasets

Muhammad Shahroz Nadeem[1,2] · **Fatih Kurugollu**[1,3] · **Sara Saravi**[4] · **Hany F. Atlam**[1] · **Virginia N. L. Franqueira**[5]

Received: 9 February 2022 / Revised: 24 June 2022 / Accepted: 22 April 2023

Abstract

Manually labelling datasets for training violence detection systems is time-consuming, expensive, and labor-intensive. Mind wandering, boredom, and short attention span can also cause labelling errors. Moreover, collecting and distributing sensitive images containing violence has ethical implications. Automation is the future for labelling sensitive image datasets. Deep labeller is a two-stage Deep Learning (DL) method that uses pre-trained DL object detection methods on MS-COCO for automatic labelling. The Deep Labeller method labels violent and nonviolent images in WVD and USI. In stage 1, WVD generates weak labels using synthetic images. In stage 2, the Deep labeller method is retrained on weak labels. USI dataset is used to test our method on real-world violence. Deep labeller generated weak and strong labels with an IoU of 0.80036 in stage 1 and 0.95 in stage 2 on the WVD. Automatically generated labels. To test our method's generalisation power, violent and nonviolent image labels on USI dataset had a mean IoU of 0.7450.

Sara Saravi, Hany F. Atlam, and Virginia N. L. Franqueira contributed equally to this work.

✉ Muhammad Shahroz Nadeem
M.Nadeem@derby.ac.uk; S.Nadeem3@uos.ac.uk

Fatih Kurugollu
FKurugollu@sharjah.ac.ae

Sara Saravi
S.Saravi@lboro.ac.uk

Hany F. Atlam
H.Atlam@derby.ac.uk

Virginia N. L. Franqueira
V.Franqueira@kent.ac.uk

[1] College of Engineering and Technology, University of Derby, Markeaton Street, Derby DE1 1DW, United Kingdom

[2] School of Engineering, Arts, Science and Technology, University of Suffolk, Ipswich, United Kingdom

[3] College of computing and informatics, University of Sharjah, Sharjah, United Arab Emirates

[4] Loughborough University, Epinal Way, Loughborough LE11 3TT, United Kingdom

[5] Department, University of Kent, Giles Ln, Canterbury CT2 7NZ, United Kingdom

Published online: 24 June 2023

Keywords Data labelling · Violence detection · WVD · USI · Synthetic virtual violence

1 Introduction

Deep Learning has made advancements in computer vision. Especially in object detection methods that use large, accurately labelled datasets to train their models, like Pascal-VOC [8] and MS-COCO [14]. Manually labelling images is a time-consuming and labor-intensive process. Further, mind wandering, boredom, and attention span affect human labelling process [11]. Therefore, data labelling solutions are expected to reach 4.1 billion by 2024, up from 1.7 billion in 2019. Naturally, Google, Amazon, and Baidu offer these expensive services [7]. When labelling sensitive image datasets, like violence, the situation worsens. Detecting violence is a high-level task due to the presence of certain visual and audio features (e.g. blood, weapon, fire, blast, screams) [5]. Such sensitive data has ethical implications, so traditional labelling processes may not be desired approach for human labelling of violent images and videos. Further this can have psychological impact on human labellers. Due to this, data corpus for violence tends to be based on movie scenes (e.g. Hollywood2 [19], Hockey Fight and Movies [24], and VSD [5]). Nadeem et al. [21] proposed an alternative way to generate violent videos to avoid ethical and psychological issues. Their Weapon Violence Dataset (WVD) is based on the open world game Grand Theft Auto-V. (GTA-V). This is the only known synthetic dataset for violent videos with weapons.

In this work, our aim is to label WVD dataset for violence detection. Specifically, our focus here is to produce a method that can automatically label person-to-person fights using off-the-self object detection methods with zero human interference. It must be noted that WVD does not previously contain any bounding box labels. Therefore, just to test our method a small subset of WVD and UNITN Social Interaction (USI) [29] datasets are manually labelled. To this end, our experiments generate bounding box labels for person-to-person fights on both synthetic and real world images. It must be emphasised that once the training process is completed the manually labelled datasets are used to report the final performance of our method. Such a method can greatly benefit the task of violence detection which is increasingly becoming important in video surveillance domain [13, 32]. Primary example include: CCTV, handheld/mounted dashcams, cellphones and drones. If such captured corpus of data can be labelled automatically, this can help create methods for Crowd Surveillance [9, 30].

Automatic labelling has been a desired aspect for supervised learning algorithms. Many such methods have been proposed for computer vision tasks. Liu et al. [16] proposed automatically labelling, detecting, and tracking football players. Xiang et al. [31] developed a method to label data for behaviour profiling and abnormality detection. Bah et al. [3] considered labelling unsupervised UAV weed images. Papadopoulos et al. [25] removed human labellers but used human verification of machine-generated labels. All such methods rely either on handcrafted features or require human involvement in labelling. However, to the best of our knowledge no such method has been proposed for automatic detection of violence. Similar to Papadopoulos et al. [25] human labelled test set is only utilised once for verification to report the final performance of the method.

To conclude, we propose a new approach that generates labels for person-to-person violent scenes without any human intervention during training. Human-generated WVD and USI test sets are only used once for verification/evaluation. The experimentation process use FRCNN [27], Yolo, Tiny-Yolo [26], RFCN [4], SSD [18], and RetinaNet [15] trained on MS-COCO [14]. It must be noted that these methods are primarily designed for object detection and not for labelling violence. However, in this paper learned low-level feature representations are used to label violence (a high level task). Instead of the bottom-up approach training DL from labelled data, our approach is top-down. The approach is two-staged. Stage 1 uses pre-trained object detection models and an aggregation function to generate weak labels for WVD dataset. The models performance is evaluated on a human-labeled set. Stage 2 retrains the existing models using the weak labels generated in stage 1 as input data. Strong labels are generated by directing the loss function to focus more on high IoU labels, then performance is again measured against the same human-labeled WVD test set. After training, the same models label a real-world dataset (USI).

The paper's contributions are:

1. Proposed a two-stage method using weak and strong learners to label virtual violence.
2. Generated human labelled test sets for WVD and USI datasets.
3. Demonstrated that DL learned feature representation (MS-COCO) can be hierarchically combined and applied directly to synthetic virtual images without retraining or transfer learning for violent scenarios (stage 1).
4. Training on weak labels (produced in stage 1) improved labelling performance, producing a strong learner that labels synthetic virtual (WVD) and real world (USI) images with high IoU scores without using temporal and spatial features (stage 2).

The remaining of the paper is organised as follows. Section 2 describes related automatic labelling approaches. Section 3 describes the virtual (WVD) and real-world (USI) datasets used in the experiments. Section 4 explains the challenges of human labelling. Section 5 describes how we generate bounding boxes for WVD and USI datasets. In Section 6, we analyse our method's WVD and USI performance. In Section 7, proposes solutions and future directions. Our findings are summarised in Section 9.

2 Related work

The concept of automatic labelling of raw data is one of the most desired aspects of supervised learning. In order to develop generic feature representation, the first step is to provide a learning algorithm with accurately labelled data. Especially for DL algorithms, these aspects hold critical value. To this end, many unsupervised and semi-supervised approaches have been developed for automatic labelling of data. Specifically, for computer vision tasks some methods have been devised for person tracking, face detection, object detection and semantic labelling, caption generation, behaviour profiling and anomaly detection. Unfortunately, not many recent attempts could be found for automatically labelling of violence, that exists in literature. This section would elaborate on the methods created to label images for vision tasks.

Xiang et al. [31] developed a method which automatically labels data for behaviour profiling and abnormality detection. Instead of object tracking, they utilised discrete scene event features which are modelled by Dynamic Bayesian Network to calculate the affinity matrix for the behaviours. Then, a Multi-Observation Hidden Markov model was used to model each behaviour and spectral clustering was performed to generate the labels. The approach is simple; however, parameter tuning for the threshold value is challenging.

Liu et al. [16] proposed a method for automatic detection, labelling and tracking of players on a football pitch. Detection is done by combining dominant color-based background subtraction and boosting detection based on Haar features. Player labelling in teams is carried out by subtracting the background from the image and converting the pixel values to the CIE-Luv color space. These pixel values are then clustered and a bag of feature representation is generated. However, when occlusion of players with indistinct appearance occurs, it results in poor performance. Nevertheless, the method is computationally inexpensive.

Recently, Bah et al. [3] considered labelling unsupervised UAV images of weed in precision agriculture. They utilise the concepts of inter-row weeds to first discriminate between the planted crops and weeds. Thus, they generated a dataset containing images for both crops and weeds. Their methodology has three basic components: 1) Line detection, 2) Inter-line weed detection, 3) Database creation and training deep network. Another unsupervised method proposed by Niebles et al. [22] used Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) to categorise and localise actions on an unseen video sequence.

Le et al. [12] trained a face detector without using labelled images. They argued that high level features can be learned from unlabelled data. Frameworks for low level feature detection are abundantly present in the literature. Similar to this, our focus in this paper is to label a high level task of violence utilising low level features. Similarly, Dong et al. [6] use face detector to label facial age of people. They trained two deep networks, the first networks is trained to detect face. The goal is to extract all the possible discriminative facial features necessary for face detection. Once this is done, transfer learning is applied to estimate age ranges for the faces.

Niemeyer et al. [23] generates semantic labels by breaking the image into perceptually coherent regions called the superpixels. These superpixels are then clustered together by Simple Linear Iterative Clustering (SLIC). Afterwards using SIFT and Hue histogram, feature representations are calculated and clustered together forming a visual bag of word dictionary. This is followed by Hierarchical Dirichlet Process (HDP) and labels are generated.

Aljundi et al. [2] proposed a method to automatically label human actor by extracting their facial features from images taken from Internet Movie Database (IMBD) without the use of subtitles or transcripts. The facial detection and description is done using Deformable Parts Model (DPM), and Deep Face Model (DFM). Afterwards face tracking is performed which results in multiple facial snaps taken from the video. This is passed to the Hungarian graph matching algorithm for clustering. Based on this the track are then assigned a label.

labelling data through man-in-the-loop is another technique utilised by Meng et al. [33] in their multi-stage weakly supervised data annotation method for 3D point cloud object detection. They predict the view angle by bin based regression method. In stage 1 they learn cylindrical 3D region proposal, stage 2 gives cascaded cuboids encircling the object.

Many such automatic labelling techniques have been designed to solve the problem of labelling huge amounts of visual data. Most of these techniques either are based on some rule based unsupervised methods where groups and labels are generated using some common patterns/features or exploit some underlining domain specific key understanding of features through which labelling is done for many vision tasks. Other techniques involve human observers to access/verify the labelling quality in the labelling process [25]. Moreover, loosely/partially labelled data or a combination of different conventional or DL based techniques are joined together to perform the labelling tasks.

However, to the best of our knowledge no such method exists for automatic labelling of violence. Mainly due to the reason, that violence related data is sensitive and is subjected to ethical implications. Furthermore, prolonged exposure to such content is discouraged. Moreover, mostly violence datasets are movie based and transcripts aid in the labelling process. Keeping these factors in mind, in this paper, a novel technique for labelling and tracking violence has been proposed using pretrained DL object detection methods. A high level tasks has been labelled by combining low-level features together in this work.

3 Datasets

The discovery of violence is frequently fraught with ethical and moral issues. Furthermore, using an open source labelling service to classify data related to violence is problematic. WVD dataset [21] was chosen for training, validation, and testing in stage 1 due to these factors. The final testing was done on the USI dataset in stage 2 to test the generalisation of our method on a real-world dataset.

The WVD dataset was created using the open world game GTA-V to create a synthetic dataset for weapon-based fights. The dataset is public and can be found on Kaggle[1]. WVD contains fights with a range of *Hot* (pistols, shotguns or automatic rifles) and *Cold* (bat, knife or broken bottles etc.) weapons. Specifically, it contains **10 hot** and **9 cold** weapon types. Each fight sequence is shot in a variety of lighting situations, including **Dusk**, **Morning**, **Midday**, **Afternoon**, **Sunset**, and **Midnight**. WVD has a stringent policy of only allowing two non-playable characters (NPCs) each fight. Only **one** of the two NPCs holds a weapon, illustrating an aggressor and victim scenario; yet, both NPCs battle with what they have. It should be emphasised, however, that both NPCs battle until one of them dies or is knocked out. However, due to the game's fight mechanics, the NPC carrying the weapon usually survives, which is typically the case in real life. Blood, gunfire flash, falling/knockout motion, varied combat stances (aggressive, regressive, and defensive), and customisable weapon usage motion are all visible visual characteristics of WVD. As a control class, there are nine different nonviolent actions between two NPCs. Yoga, gardening, dancing, exercising, talking, argumentation, construction, vehicle maintenance, and being arrested are all examples of this.

On the other hand, USI [2] dataset [29] is essentially a social interaction dataset that contains four human interactions: **Talking**, **Shaking**, **Hugging**, and **Fighting**. Three of these interactions are nonviolent, whereas only cold hand-to-hand violence is included in the fighting group. It should be emphasised that the USI dataset contains no hot fights. However, similar to WVD, this dataset covers interactions between up to two people, with a total of 16 variable-length movies per interaction. Videos from the USI dataset are also collected in a frontal or dash-cam view, similar to WVD. It should be mentioned that the embracing and fighting classes are the most difficult to distinguish.

4 Human labels generation: WVD and USI dataset

The development of human labels to evaluate the labels created for the WVD and USI was an important component of our technique and an essential aspect of our testing procedure. To this end, we randomly selected 10% of photos from the WVD for human labelling as the test set for evaluation purposes. The validation set comprised 20% of the total dataset.

[1] https://www.kaggle.com/thelarka/weapon-violence-dataset-wvd.

[2] http://loki.disi.unitn.it/USID/.

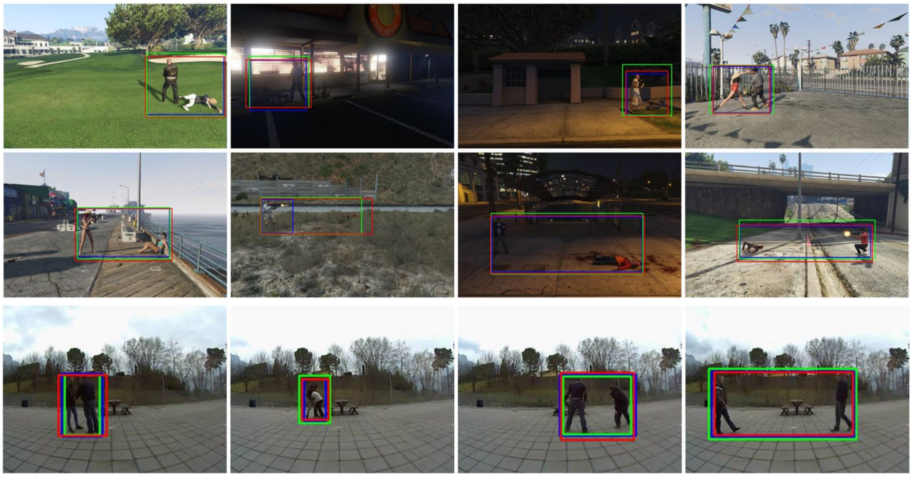

Fig. 1 Bounding Boxes generated by different human labellers for the same images of WVD and USI datasets

From a total of 40,845 photos, this yielded 4,085 frames for the test set and 8,169 frames for validation. After that, 15 people were randomly assigned to the test set for human labelling. Our basic rule was to just look for fights, and each participant was instructed to create bounding boxes around the Region of Interest (RoI). Human labellers were in charge of all other aspects.

A few intriguing insights about the participants' labelling behaviour have been discovered. For the same set, each participant came up with radically different bounding boxes. As a result, the bounding box test set coordinates differed. This raises the question of which test set should be taken into account and why. Furthermore, what significance would the results provided on each test set have? To tackle this problem and account for human labelling variances, the test set was divided into 6 (WVD) and 5 (USI) equal halves. By integrating tagged chunks from the labelled test sets, a single test set was created. The inclusion of splattered blood, shadows, and clothing items (such as hats or glasses) as part of the RoI, while others ignored these visual characteristics entirely, is the basis for labelling variations. Furthermore, human labellers found it particularly challenging to correctly label violent sequences in low-light situations, which exacerbated labelling mistakes. These findings imply that humans' understanding of a violent scenario and the production of bounding box coordinates are highly subjective.

The variations in human labelling of the WVD are shown in Fig. 1 for clarity, just a sample of three bounding boxes out of the total of six for WVD is shown. Each different coloured bounding box depicts a human-drawn bounding box on what they regarded as violence.

5 Proposed approach

The identification of violence is a difficult process since it necessitates the presence of specific visible and audible elements such as blood, fire, weapons, screaming, or explosions. Furthermore, motion and temporal features are frequently used. The goal

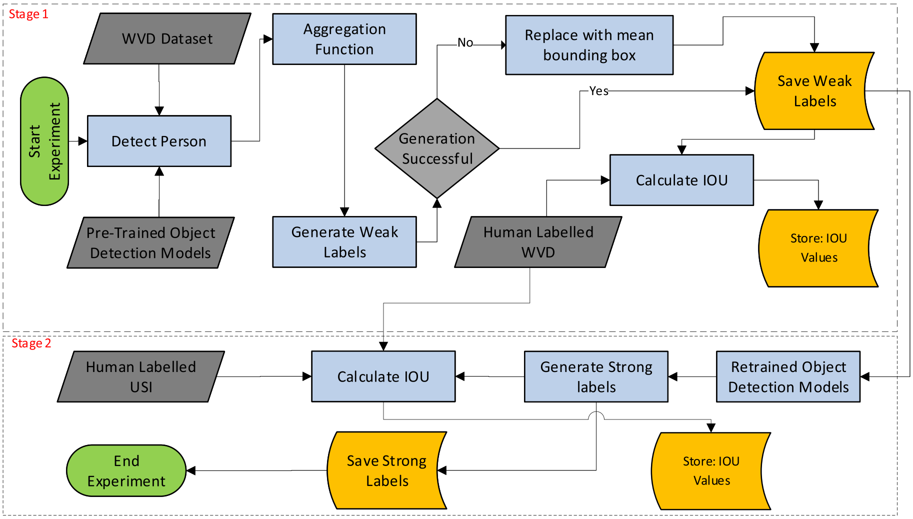

Fig. 2 Flow chart of the proposed two-stage automatic label generation method

of this project is to construct bounding boxes for a (high level) violence detection problem using simple (low level) object detection algorithms. We suggest that a high-level knowledge of events necessitates the hierarchical organisation of low-level features. With this in mind, an automatic approach is designed that generates weak and strong labels, reducing human interaction in the labelling process to a minimum. There are two stages to the proposed technique.

1. **Stage 1**: In this stage, we use object detection algorithms that have been pre-trained to generate weak labels by identifying and aggregating (Virtual NPCs) people in the frames. After the bounding boxes have been aggregated, the labelling quality is determined by comparing the IoU of the bounding boxes to human-labeled WVD frames.
2. **Stage 2**: State-of-the-art object detection techniques are retrained to produce strong labels using the weak labels generated in stage 1. WVD [21] and USI human labelled test sets are used to evaluate the strong labels.

The labels created are explained in detail below; it should be emphasised that these labels are generated without human interaction; only a subset of the datasets is labelled by humans for final evaluation of the approach.

Weak labels. *are built automatically by combining the boundaries generated by object detection algorithms trained on MS-COCO with the aggregation function. These are labels with both high and low IoU values.*

Strong labels. *are generated by retraining the object detectors on weak labels by directing the loss function to focus on only labels with high IoU values.*

The two-stage technique for creating labels is shown in Fig. 2. In step 1, the images depict the process of creating weak labels in detail. Persons are first discovered using pre-existing object detection methods. As described in equation 2, the created bounding boxes are then supplied to the aggregate function. The aggregate function's objective is to build a broader bounding box that encompasses the entire area of violence. The bounding box values are saved if the generation is successful. In the absence of this, a mean bounding box value is calculated, generated and saved. It must be noted that

Meanwhile, stage 2 shows strong labels generated through training a strong learner. It must be noted that in stage 1 all the learning of DL methods has been carried out using the WVD dataset. Here the most commonly known state-of-the-art pre-trained object detection methods are used to detect persons in the frames. It is worth mentioning that the DL methods used in stage 1 are trained on real world images of persons. However, in this stage, we apply the learned features representation of MS-COCO directly to virtual persons in WVD. Once the bounding boxes are generated, they are passed through the aggregation function.

The produced labels are then checked which we refer to as weak labels. This process is error prone due to false positives, and misclassifications due to occlusion and deformation. In case of misclassifcation we replace the bounding box with a mean bounding box value taken from the previous three bounding boxes. However, this process just replaces missing bounding boxes and actually has a negative impact on IoU performance of stage 1. Due to this performance degradation in stage 1, weak labels are produced. Once this process is complete, the weak labels are evaluated against a human labelled test set.

In order to balance this negative impact of misclassification and false positive in stage 1. In stage 2 the weak labels are then fed to the object detection methods, the DL methods are retrained and networks are directed to focus more on labels which produce good IoU scores. The goal here is to make the DL methods focus more on the good labels rather than the labels produced due to inherent weaknesses in stage 1. Through this process strong labels are generated. These labels are then evaluated against the WVD test set. Once the training, validation and testing phases are completed, a final pass is performed on the USI dataset.

5.1 Evaluation criterion

The localization and classification evaluation of the item can be divided into performance analysis of object detection methods. The most often utilised evaluation matrices for these purpose are Mean Average Precision (mAP) [17] and IoU. However, we are just interested in evaluating the localization of violence in this study. IoU is the de facto evaluation metric for evaluating the model's localisation performance [28]. As demonstrated in Eq 1, IoU (also known as Jaccard Index) divides the area of intersection by the area of union for the overlapping bounding boxes. The IoU metric produces a normalised value between '0' and '1,' with '0' indicating no overlap and '1' indicating total overlap. For photos with both human and automatic machine labelling, IoU values are calculated. This procedure is carried out for each of the pre-trained object detection algorithms used in this study.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

5.2 Implementation details

We employ the Tensorflow Object Detection API [10] and ImageAI [20] libraries to construct the machine labels for violence detection because our approach uses pre-trained object detection algorithms. Only the implementations with the highest detection rates were chosen from each library. Because the purpose is to develop labels for violence, only the most effective models were chosen. It should be highlighted that this paper does not present a novel model for detecting violence. Our approach, on the other hand, is able to create high IoU values (i.e. localization refer back to Section 5.1) through automatic labelling, resulting in high detection scores.

Tiny-Yolo and **Yolo** [26], **Faster-RCNN (FRCNN)** [27], **RFCN** [4], **SSD** [18] and **RetinaNet** [15] were among the deep networks employed in this study. All of these algorithms have been pre-trained on the **MS-COCO** dataset [14], a large-scale dataset with 80 different types of objects. The "Person" class is the sole object utilised in this method. Table 1 shows the hyper parameter values set in the experiments. The people in the WVD are detected using these pre-trained networks. As previously stated, everyone in the WVD are essentially NPCs from the photo-realistic game GTA-V. Consequently, despite the fact that these networks were trained to detect humans using photos of real-world people, they are used unchanged in our technique.

When a person involved in a conflict is identified, their bounding boxes are used as anchors to create broader bounding boxes for violence. Four values (x, y, w, h) are returned for each successful detection, where x and y are the coordinates of a point on the image, and w and h are the width and height (respectively) of the bounding box with respect to the point (x, y) being the top-left corner of the bounding box.

The bounding boxes created by the pre-trained object detection network (for person detection) are first aggregated to localise the RoI for the entire conflict. Eq 2 describes the aggregation function that was used. These bounding boxes are then stored, and a record is kept for each violent WVD scenario. When the object detectors fail to detect the person, these records are used to replace bounding boxes. As the WVD battle sequences feature violence, they were captured with a variety of lighting settings, as well as NPC occlusion, deformation, and size modifications (see Figs. 1, 6, 7, and 8). Furthermore, because object detectors are dependent on real-world data, using them in a virtual environment can result in a high number of false positives and misclassifications. Previous bounding boxes are recorded and used instead of a mean bounding box value to address detection misses. In order to achieve this, the previous three saved bounding boxes are used to produce a mean bounding box value, which is then replaced when the object detector fails (cf Fig. 2). The (semi accurate) weak labels and the IoU value for individual photos and the dataset as a whole are saved after the process is complete. The first stage has come to a conclusion.

$$\begin{aligned} f_{agg} = (&\min\left(x_1, x_2\right), \\ &\min\left(y_1, y_2\right), \\ &\max\left(x_1 + w_1, x_2 + w_2\right), \\ &\max\left(y_1 + h_1, y_2 + h_2\right)) \end{aligned} \tag{2}$$

The models were retrained using the weak labels created in stage 2 using the Google object detection API. The 0.0001 learning rate and early stopping method resulted in superior IoU values at this stage. The early halting compels the network to only learn the good labels because they are the majority, as shown in Fig. 3, while disregarding the negative labels that resulted from the false positive in stage 1, as shown in Fig. 3. As a result, getting high IoU scores at the end of stage 2 requires a modest learning rate and an early halting approach.

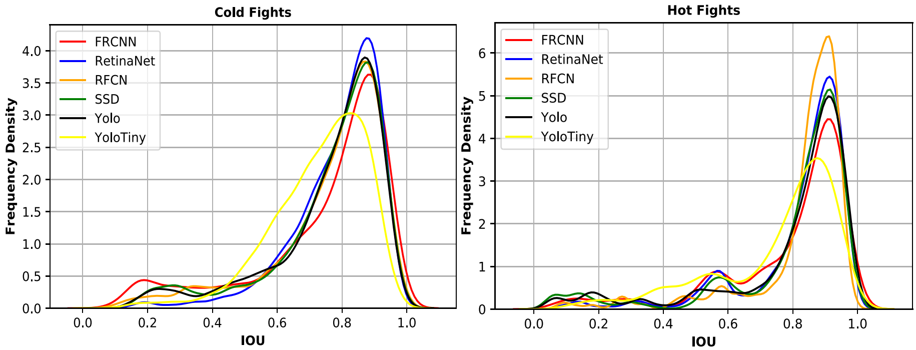

Fig. 3 Comparison of frequency density against IoU values for each object detection method

Table 1 The set of hyper parameters utilised in the experimentation process

Hyper Parameters	Values
Learning Rate	$1\epsilon^{-2}$ - $1\epsilon^{-6}$
Gradient Descent Optimiser	ADAM
Training Epochs	10 - 100
Training set WVD	72286
Validation set WVD	10326
Test set WVD	20654

6 Results

Machine labels were created for all of the item detection methods stated in Section 5.2 once the human labeling process was done. The performance of each object detection algorithm against each weapon type in the WVD is shown in Figs. 4 and 5. Object detection algorithms **FRCNN**, **SSD**, **Yolo**, and **Tiny-Yolo** fought for the “Golf Club and Knife” weapon categories in the cold weapon category. Overall, **RetinaNet** functioned admirably. For **SSD** and **Tiny-Yolo**, however, the most difficult weapon category was “MGCombat,” whereas **FRCNN** only struggled with the “RifleCarbine” category. The performance of **RetinaNet** and **RFCN** appears to be somewhat similar (Fig. 5). Because each scenario in the WVD has distinct scale variations (such as NPC occlusion and distortion), each weapon’s IoU value varies depending on the time setting.

Figure 3 illustrates the frequency distribution of the labels produced for violence to have a better idea of the performance of object detectors. Both graphs for hot and cold fights are clearly skewed, indicating that the majority of the labels generated have a higher IoU value. Hot labels, on the other hand, have a better probability of receiving a higher IoU value in our method. **RetinaNet** and **RFCN** have proved their supremacy in virtual settings for cold and hot conflicts, respectively, in Fig. 3.

The total performance of each object detection for the whole WVD is shown in Table 2. **RFCN** had an overall IoU value of **0.8256** for hot fights, while **RetinaNet** had an IoU value of **0.7882** for hot fights. Overall, **RetinaNet** had the highest IoU value of **0.8036** for the entire dataset. This demonstrates that our technology is capable of automatically labelling synthetic data like the WVD dataset. This would allow for the creation of large-scale

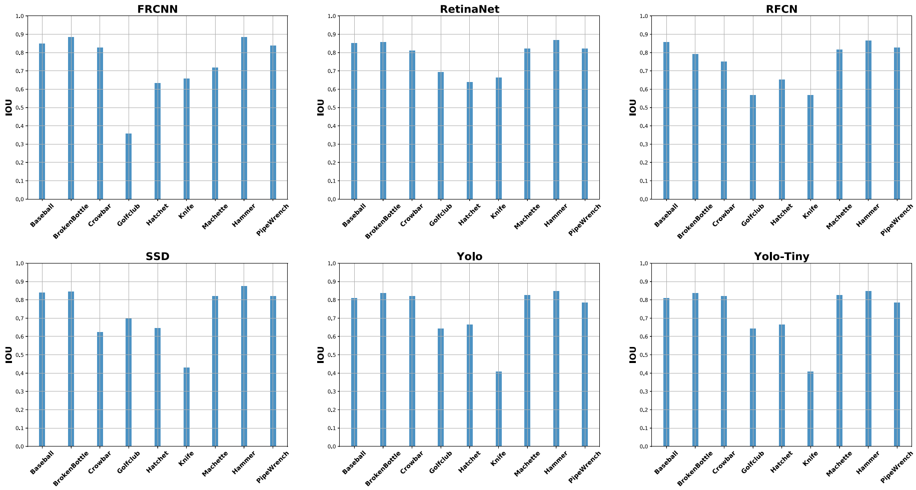

Fig. 4 Overview of the achieved IoU values for each object detection methods with respect to cold weapon type fight sequence in the WVD dataset

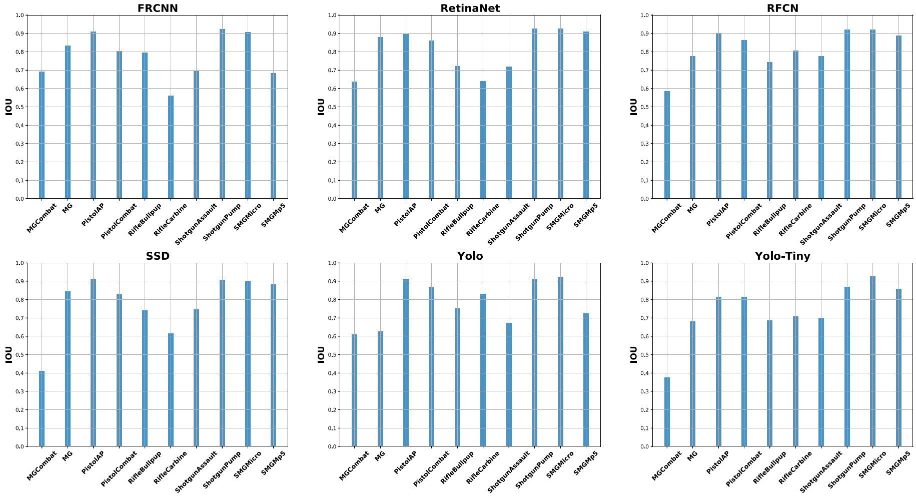

Fig. 5 Overview of the achieved IoU values for each object detection methods with respect to hot weapon type fight sequence in the WVD dataset

Table 2 Summary of achieved IoU values for each object detection method on WVD

	FRCNN	RetinaNet	RFCN	SSD	Yolo	Tiny-Yolo
Hot	0.7864	0.8190	0.8256	0.7939	0.7925	0.7513
Cold	0.7440	0.7882	0.7607	0.7605	0.7596	0.7310
Total	0.7652	0.8036	0.7932	0.7772	0.7761	0.7412

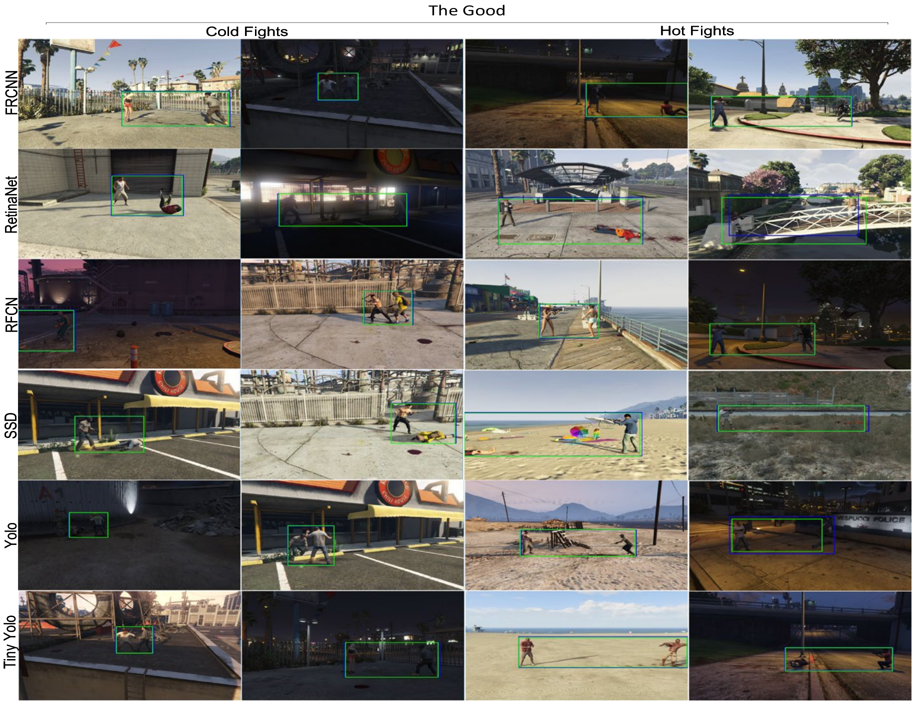

Fig. 6 Examples of "The Good" labels: the green bounding boxes represent machine labels while blue represents the human labels for violence detection

databases for sensitive sectors like violence, lowering human exposure and ethical considerations while labelling.

6.1 The Good, The Bad, and The Ugly labels

After reaching a high IoU score using our automatic labelling approach, the achieved performance is analysed, and the labels are classified into three categories based on the IoU scores: "The Good", "The Bad," and "The Ugly." Images having an IoU value more than **0.8** are referred to as **"The Good"**, while images with an IoU value between **0.5 and 0.8** are referred to as **"The Bad"**, and images with an IoU value less than **0.5** are referred to as **"The Ugly"**. However, it must be noted that in object detection IoU values greater than 0.5 are considered good.

Our technique produces "The Good" labels, which are compared to human labels (Section 4) in Fig. 6. The latter is represented by the blue bounding boxes, whereas the former is represented by the green bounding boxes. Human and machine-generated labels are nearly similar and entirely overlapping, as seen in the diagram. Furthermore, the WVD's performance has been stable across all time settings.

"The Bad" designations are shown in Fig. 7. In most cases, an IoU value greater than 0.5 is considered *good* in object detection. In the instance of violence detection, however, IoU values of approximately 0.5, in our opinion, cannot be considered good. The reason for this is that in violent situations, the chance of death is higher. The figure shows how human

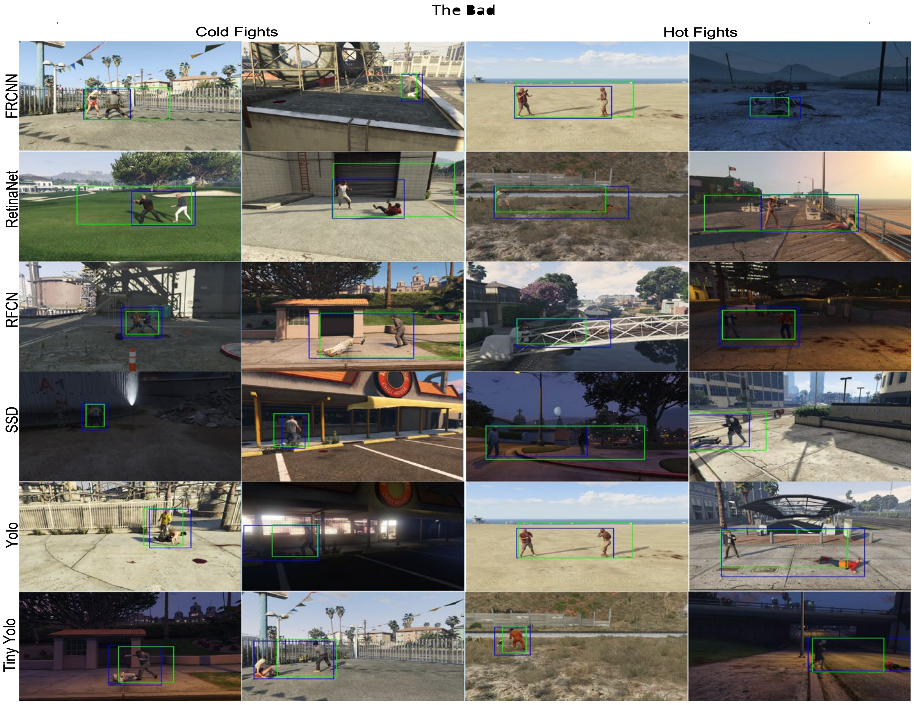

Fig. 7 Examples of "The Bad" labels: the green bounding boxes represent machine labels while blue represent the human labels for violence detection

and machine labels partially overlap; while there is some violence here, the machine labels are covering extra ground in the majority of cases. In the case of the "The Ugly" labels, as illustrated in Fig. 8, there is essentially no or very little overlap.

Poor performance can be attributable to three variables after a closer examination of the photos. First, the false positive ratio has an effect on the aggregation function (Eq 2). The object detection algorithms were trained to recognise humans from real-world data (MS-COCO), but we tested them on synthetic data captured in virtual environments with no preprocessing. This explains why object detectors misclassified certain things as *person* due to differences in learning feature representation. Figures 7 and 8 illustrate that some false positives were found, as seen by the expanded machine produced bounding boxes.

Second, after a successful construction of the violence bounding box, the coordinates are retained in order to anticipate the individual in the event of a failed detection. The irregularities are caused by the misleading positive violent bounding boxes. It's also worth noting that smaller records in the past result in poor forecasts for the first frames in the series. This suggests that rather than relying on previously recorded bounding boxes, an intelligent approach is needed to forecast the current violent bounding box.

Third, the object detectors completely miss one of the fight participants due to NPC occlusion and deformation in consecutive frames. Object detectors are almost certain to fail on NPCs that have fallen to the ground, are out of the capture frame, or are occluding, as shown in Fig. 8. Images of RetinaNet and YOLO, specifically for hot fights, are an excellent example of this trend. Cold fight labels for Tiny-Yolo, SSD, and RetinaNet

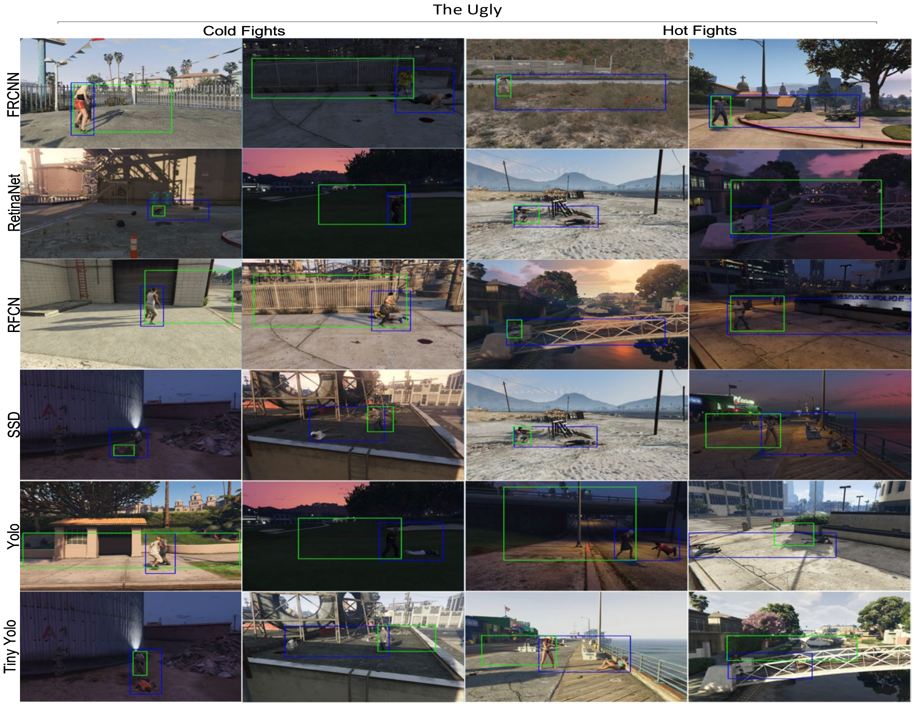

Fig. 8 Examples of "The Ugly" labels: the green bounding boxes represent machine labels while blue represents the human labels for violence detection

are also affected by poor lighting and NPC occlusion. The fact that a fallen NPC is frequently missed by object detection methods can be seen here. Performance is also affected by occlusion with the background and deformation issues.

These characteristics have an impact on the repository of previously recorded bounding boxes, making it difficult to anticipate the location of a missing individual and thus prone to mistake propagation. Despite these difficulties, our system provided high-quality labels and overall performance.

6.2 Performance on USI test set

After completing stage 1, all of the labels were mixed at random and the object identification techniques were retrained to produce strong labels. Early stopping approach was used during the retraining process so that the system focuses more on the good labels rather than the bad or ugly ones. With a value of **0.95**, the FRCNN gave the highest IoU score for the WVD validation set after retraining. Once strong labels and learners are created, their performance is evaluated on a real-world dataset in the end. For this purpose the USI human-labeled test set was run via the procedure that produced the most IoU. (i.e. FRCNN). Table 3 lists the performance that was recorded. On the USI dataset, the final IoU score was **0.7450**. Figure 9 shows "The Good", "The Bad", and "The Ugly" labels for the USI datasets. As with WVD, the labels generated here cover the majority of the ROI for violence. Even the Bad labels cover the majority of the region where individual exchanges

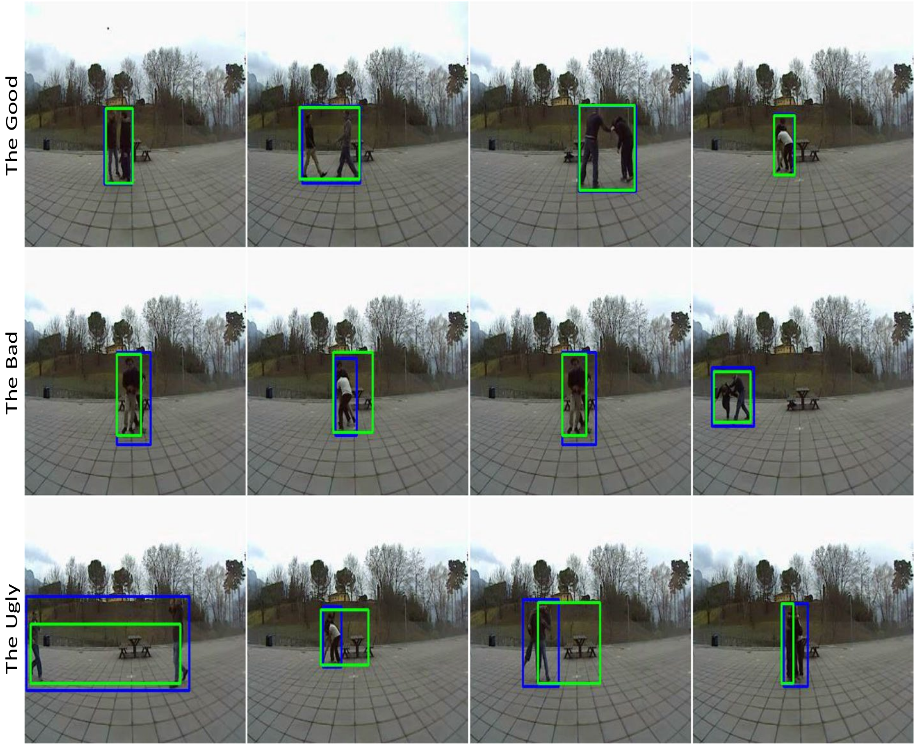

Fig. 9 In the figure green bounding boxes represent machine labels while blue represent the human labels on the USI dataset

Table 3 The table shows the overall performance achieved on WVD (stage 1 and 2) and USI dataset

	IoU Score
Stage 1 (WVD) (RetinaNet)	0.8036
Stage 2 (WVD) (FRCNN)	0.9500
USI dataset (FRCNN)	0.7450

take place. This demonstrates the method's robustness, demonstrating that it can function for both virtual synthetic and real-world photos with minimal human interaction. It's worth noting that all of the training took place on the virtual WVD dataset.

7 Future work

Our approach to label creation is applicable to one-on-one fights; however, expanding this strategy to crowd fights might be a fascinating path to pursue. Another potential avenue is the detection of person-to-person fighting in a virtual crowd scenario by training object detectors on the generated labels. Human labelling suffers from human psychological concerns such as mind wandering or boredom, as demonstrated in our experimentation approach. These human characteristics can introduce inaccuracies and differences, resulting in distinct bounding box coordinates. Deep labellers would tackle this problem by generating a large amount of labelled data in a short amount of time, all with the same bounding box co-ordinates. Low-level features could be used to create hierarchical representations for labelling high-level tasks in other computer vision applications as well. However, such technologies might be tailored to specific areas, and automatic data labelling with an acceptable error rate can still save time and effort.

Labels with high IoU values were made using the proposed method. However, depending on the mean bounding box value determined from previously created bounding boxes is not a robust/smart method in the case of poor detection from the pre-trained networks. Furthermore, inaccuracies and the quality of poor labelling have a significant impact on performance. One alternative option is to use combat motion data to estimate where the NPCs will appear. Additionally, recurrent networks have been shown to boost performance ([1]).

8 Limitations of our approach

Despite the fact that our method yields high IoU labels for WVD and USI datasets, it is not without flaws. We must clarify that the approach used in this study is intended for simple person-to-person battles. This is done on purpose to test our hypothesis of using DL object detection technologies to categorise violent videos from the top down. This research demonstrates that synthetic movies for violence can be labelled with high-quality labels without the need for human intervention. This strategy, however, cannot be immediately applied to situations involving crowd violence. Furthermore, the performance of the off-self object identification methods used in our methodology can have an impact on overall labelling performance. As a result, we tested a variety of cutting-edge object detection techniques.

9 Conclusion

Using object detection models pre-trained on the MS-COCO dataset, we developed a method for automatically constructing bounding boxes for violence detection in a virtual environment. A subset of the Weapon Violence Dataset (WVD) and the UNITN Social Interaction (USI) datasets were labelled by 15 human volunteers as part of our experimental procedure. These test sets will be used in our research and are publically available on Kaggle.

Our method for automatic labelling demonstrated that low-level information extracted and combined from simple object detectors may be utilised to label high-level tasks like violence detection. As a result, it was discovered that Deep Learning-based object detection methods might be used as a **deep labeller** to label a huge number of photos with better IoU performance. The method can be used to classify sensitive photos with little or no human intervention and exposure. Furthermore, our research revealed that Deep Learning algorithms trained on real-world photographs may be applied to data gathered in a virtual environment with success.

In stage 1, RetinaNet was able to produce overall high quality weak labels with an IoU value of **0.8036** on the WVD, while RFCN and RetinaNet produced the highest IoU values of **0.8256** and **0.7882** for hot and cold fights, respectively. For all of the DL approaches utilised in our study, the overall trend for bounding box generation was positively skewed towards larger IoU values. This performance was improved to **0.95** in stage 2 by passing the weak labels learned in stage 1. Finally, a single pass of the USI test set was completed, yielding a **0.745** IoU score. This demonstrates that our method can accurately label images for violence detection with minimal human intervention.

Acknowledgements We would like to thank the following individuals for labelling violent images: Muhammad Kazim, Syed Ali Zahir Bukhari, Yasir Adnan, Jawwad Ahmad, Haider Ali, Mohannad Ali Sultan, Mohammad Sadiq Hassan zadas, Mohammad Ali, Muhammad Aamir, Zawar Nadeem, Muhammad Irmaghan Nadeem, Uwais Habib, Aleena Nadeem, Wahab Nadeem and Junaid Nadeem.

Data availability The dataset is publicly available on Kaggle.

Declarations

Conflict of interest The preprint of this paper is available on TechRXiv.

References

1. Aktı Ş, Tataroğlu GA, Ekenel HK (2019) Vision-based fight detection from surveillance cameras. In: 2019 ninth international conference on image processing theory, tools and applications (IPTA), IEEE, pp 1–6
2. Aljundi R, Chakravarty P, Tuytelaars T (2016) Who's that actor? automatic labelling of actors in tv series starting from imdb images. In: Asian conference on computer vision, Springer, pp 467–483
3. Bah MD, Hafiane A, Canals R (2018) Deep learning with unsupervised data labeling for weed detection in line crops in uav images. Remote Sens 10(11):1690
4. Dai J, Li Y, He K, et al (2016) R-fcn: Object detection via region-based fully convolutional networks. In: Proceedings of the 30th international conference on neural information processing systems. NIPS'16, Curran Associates Inc., Red Hook, pp 379–387
5. Demarty C, Ionescu B, Jiang Y, et al (2014) Benchmarking violent scenes detection in movies. In: 2014 12th international workshop on content-based multimedia indexing (CBMI), pp 1–6
6. Dong Y, Liu Y, Lian S (2016) Automatic age estimation based on deep learning algorithm. Neurocomputing 187:4–10
7. Eastwood JD, Frischen A, Fenske MJ et al (2012) The unengaged mind: defining boredom in terms of attention. Perspect Psychol Sci 7(5):482–495
8. Everingham M, Van Gool L, Williams CK et al (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
9. Fradi H, Luvison B, Pham QC (2016) Crowd behavior analysis using local mid-level visual descriptors. IEEE Trans Circuits Syst Video Technol 27(3):589–602
10. Huang J, Rathod V, Sun C, et al (2017) Speed/accuracy trade-offs for modern convolutional object detectors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7310–7311
11. Killingsworth MA, Gilbert DT (2010) A wandering mind is an unhappy mind. Science 330(6006):932–932
12. Le QV (2013) Building high-level features using large scale unsupervised learning. In: 2013 IEEE international conference on acoustics, speech and signal processing, IEEE, pp 8595–8598
13. Li T, Chang H, Wang M et al (2014) Crowded scene analysis: a survey. IEEE Trans Circuits Syst Video Technol 25(3):367–386

14. Lin TY, Maire M, Belongie S, et al (2014) Microsoft coco: common objects in context. In: European conference on computer vision, Springer, pp 740–755
15. Lin TY, Goyal P, Girshick R, et al (2017) Focal loss for dense object detection. In: The IEEE international conference on computer vision (ICCV)
16. Liu J, Tong X, Li W et al (2009) Automatic player detection, labeling and tracking in broadcast soccer video. Pattern Recogn Lett 30(2):103–113
17. Liu L, Özsu MT (eds) (2009) Mean average precision, Springer US, Boston, MA, pp 1703–1703. https://doi.org/10.1007/978-0-387-39940-9_3032
18. Liu W, Anguelov D, Erhan D, et al (2016) Ssd: single shot multibox detector. In: European conference on computer vision, Springer, pp 21–37
19. Marszałek M, Laptev I, Schmid C (2009) Actions in context. In: CVPR 2009-IEEE conference on computer vision & pattern recognition, IEEE computer society, pp 2929–2936
20. Moses, Olafenwa, J (2018) Imageai, an open source python library built to empower developers to build applications and systems with self-contained computer vision capabilities. https://github.com/OlafenwaMoses/ImageAI . Accessed 25 June 2019.
21. Nadeem MS, Franqueira VN, Kurugollu F, et al (2019) Wvd: A new synthetic dataset for video-based violence detection. In: International conference on innovative techniques and applications of artificial intelligence, Springer, pp 158–164
22. Niebles JC, Wang H, Fei-Fei L (2008) Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vis 79(3):299–318
23. Niemeyer M, Arandjelović O (2018) Automatic semantic labelling of images by their content using non-parametric bayesian machine learning and image search using synthetically generated image collages. In: 2018 IEEE 5th international conference on data science and advanced analytics (DSAA), IEEE, pp 160–168
24. Nievas EB, Suarez OD, García GB, et al (2011) Violence detection in video using computer vision techniques. In: International conference on computer analysis of images and patterns, Springer, pp 332–339
25. Papadopoulos DP, Uijlings JR, Keller F, et al (2016) We don't need no bounding-boxes: training object class detectors using only human verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 854–863
26. Redmon J, Divvala S, Girshick R, et al (2016) You only look once: Unified, real-time object detection. In: The IEEE conference on computer vision and pattern recognition (CVPR)
27. Ren S, He K, Girshick R, et al (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Cortes C, Lawrence ND, Lee DD, et al (eds) Advances in neural information processing systems 28. Curran Associates, Inc., pp 91–99
28. Rezatofighi H, Tsoi N, Gwak J, et al (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: The IEEE conference on computer vision and pattern recognition (CVPR)
29. Rota P, Conci N, Sebe N (2012) Real time detection of social interactions in surveillance video. Computer vision-ECCV 2012. Springer, Workshops and demonstrations, pp 111–120
30. Wang T, Qiao M, Lin Z et al (2018) Generative neural networks for anomaly detection in crowded scenes. IEEE Trans Inf Forensics Secur 14(5):1390–1399
31. Xiang T, Gong S (2005) Video behaviour profiling and abnormality detection without manual labelling. In: Tenth IEEE international conference on computer vision (ICCV'05), vol 1. IEEE, pp 1238–1245
32. Zhang T, Jia W, He X et al (2016) Discriminative dictionary learning with motion weber local descriptor for violence detection. IEEE Trans Circuits Syst Video Technol 27(3):696–709
33. Zhou T, Li L, Li X et al (2021) Group-wise learning for weakly supervised semantic segmentation. IEEE Trans Image Process 31:799–811