Deep Labeller: Automatic Bounding Box Generation for Synthetic Violence Detection Datasets

Muhammad Shahroz Nadeem, Fatih Kurugollu, *Senior Member, IEEE* Sara Saravi, Hany F. Atlam, Virginia N. L. Franqueira *Member, IEEE*

Abstract—Manual labelling of datasets used in training violence detection systems is time consuming, costly, and require a huge pool of human resources. Certain psychological aspects (e.g., mind wandering, boredom and reduced attention span) may also result in labelling variations and errors. Furthermore, dealing with sensitive images containing violence carries ethical implications which make their collection and distribution challenging. Therefore, we perceive automation as the way forward for labelling datasets containing sensitive images. To this end, we devise a two-stage Deep Learning (DL) method, called Deep labeller, that utilises existing pre-trained DL object detection methods on MS-COCO for automatic labelling. The Deep Labeller method is applied to the WVD and UNITN Social Interaction (USI) dataset to label violent and non violent images. In stage-1 synthetic images in WVD are used to generate weak labels. Strong labels in stage 2 are generated by retraining the Deep labeller method on the weak labels. Finally, to test the performance of our method on real world violence, USI dataset is utilised. Experimental evaluation shows that Deep labeller generated weak and strong labels with a mean Intersection Over Union (IoU) value of 0.80036 in stage 1 and 0.95 in stage 2 on the WVD. These labels were generated without any human intervention. To test the generalisation power of our method, labels generated for violent and non violent images on USI dataset had a mean IoU value of 0.7450.

Index Terms—Data labelling, Violence Detection, WVD, USI, Synthetic Virtual Violence

I. INTRODUCTION

DEEP Learning has achieved great strides in many computer vision tasks, especially in object detection. Object detection methods rely on large scale and accurately labelled datasets such as Pascal-VOC [1] and MS-COCO [2] to train their models. However, generating labelled data is a cumbersome task that requires huge human effort and intervention. Data labelling solutions market value for 2019 was estimated to be 1.7 billion dollars, and is expected to increase to 4.1 billion dollars by 2024 [3]. Naturally, such services provided by Google, Amazon and Baidu are very costly. Furthermore, mind wandering, boredom and reduced attention span affect the performance of human labellers [4], [5]. When labelling

data is required for sensitive image datasets, such as violence, the situation is worse. Violence is a very subjective concept; in contrast to normal objects, detecting violence is considered a high level task, due to the presence of certain visual and audio features (e.g. blood, weapon, fire, blast, screams) [6]. Such sensitive data is also subject to ethical implications thus applying traditional labelling processes is a challenging task. Due to this reason, violence detection datasets available tend to be based on a collection of violent scenes from movies (e.g. Hollywood2 [7], Hockey Fight and Movies [8] and VSD [6]). An alternative to avoid such ethical issues is the Weapon Violence Dataset (WVD) [9] based on the open world game Grand Theft Auto-V (GTA-V). To the best of our knowledge, this is the available synthetic dataset for violence detection.

Many methods for automatically labelling have been proposed for computer vision tasks. Liu et al. [10] proposed a method for automatic labelling, detection and tracking of players on a football pitch. Xiang et al. [11] developed a method which automatically labels data for behaviour profiling and abnormality detection. Bah et al. [12] considered labelling unsupervised UAV images of weed in precision agriculture. Papadopoulos et al. [13] removed human labellers but still relied on active human verification of machine generated labels. All such methods are either based on hand crafted features or requires some form of human role in the labelling process.

In this paper, we propose a new approach which generates labels for person-to-person violent scenes which is completely free of any human verification/intervention during the training process and features are learned automatically. Only for evaluation purposes human generated test sets for WVD and UNITN SocialInteraction (USI) dataset are utilised once the method has been trained. The DL models utilised in the experimentation process include **FRCNN** [14], **Yolo** and **Tiny-Yolo** [15], **RFCN** [16], **SSD** [17] and **RetinaNet** [18] trained on **MS-COCO** [2]. These methods are traditionally used for object detection, not labelling. We utilise learned low level feature representation and apply it to label a high level task of violence detection.

The proposed approach is divided into two stages. Stage 1 utilises existing pre-trained object detection models in combination with an aggregation function to generate weak labels for the WVD dataset. Further, the labels are evaluated on the human labelled test set. Stage 2 utilises the weak labels to retrain the existing models to generate strong labels. Once the

Muhammad Shahroz Nadeem, Fatih Kurugollu and Hany F. Atlam are with the College of Engineering and Technology at the University of Derby, Derby, United Kingdom. Email: (M.Nadeem@derby.ac.uk, F.Kurugollu@derby.ac.uk, H.Atlam@derby.ac.uk)

Sara Saravi is with the Loughborough University, Loughborough, United Kingdom. Email: S.Saravi@lboro.ac.uk.

Virginia N. L. Franqueira is with the School of Computing at University of Kent, Canterbury, United Kingdom. Email: v.franqueira@kent.ac.uk.

strong labels are generated the performance is evaluated first on human labelled test set for WVD to measure the increase in performance. Afterwards, the same trained models are utilised to generate labels for a real life dataset (USI).

The main contributions of the paper are as follows:

- 1) Proposed a two-stage method which utilises the existing DL method as weak and strong learners to label violence in a synthetic virtual environment.
- 2) Generated human labelled test sets for WVD and USI datasets.
- 3) Demonstrated empirically 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 by achieving a high Intersection over Union (IoU) score (stage 1).
- 4) Improved labelling performance through training on weak labels (produced in stage 1), producing a strong learner that labels synthetic virtual (WVD), and the 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 II gives a brief overview of related approaches to automatic labelling. Section III describes the virtual (WVD) and real world (USI) datasets utilised in the experimentation process in this paper. Section IV explains the challenges of human labelling. Section V presents our approach for the generation of bounding boxes around violence-related Region of Interest (RoI) for the WVD and USI dataset. In Section VI, we analyse the achieved performance of our method on the WVD and USI datasets. In Section VII, several solutions and promising future direction are proposed. Finally, Section VIII concludes the paper with our findings, highlighting our main contributions.

II. 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.

Liu et al. [10] 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.

Xiang et al. [11] 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.

Recently, Bah et al. [12] 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. [19] used Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) to categorise and localise actions on an unseen video sequence.

Le et al. [20] 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.

In literature such automatic labelling is mostly achieved through some sort of unsupervised technique, which is rule based, it groups and labels data using some patterns or common features which are sensitive to outliers or noise in data. In many cases, human observers are kept to access/verify the labelling quality in the labelling process [13]. However, such techniques are often not robust and for sensitive data such as violence prolong human exposure during training and evaluation may not be desirable. Keeping these factors in mind, our aim is to design a method where, human exposure is very limited. Instead of unsupervised method, DL methods with high quality low level features extraction capability is utilised to label a high level task of violence. Further, as DL methods are inherently robust methods, therefore reduces the adverse effects of noise in data.

III. DATASETS

Violence detection is often subjected to ethical and moral implications. Further, labelling violence related data through open source labelling service is also difficult. Due to these factors, WVD dataset [9] was selected for training, validation and testing in stage 1. To test the generalisation of our method on a real world dataset, the final testing was performed on the USI dataset in stage 2.

The WVD dataset is a synthetic dataset for weapon based fights built using the open world game, GTA-V. 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 performed under different lighting conditions which include **Dusk, Morning, Midday, Afternoon, Sunset** and **Midnight**. WVD has a strictly followed two Non Playable Character (NPC)s per fight policy. Amongst the **two** NPCs, only one carries a weapon depicting aggressor and victim

scenario, still both NPCs fight with what they have at their disposal. However, it must be noted that both NPC fight until one drops dead/knocked out. However, as per the fight mechanics of the game mostly the NPC with the weapon survives, which is often the case in the real world as well. Visible visual markers of WVD include blood, gunshot flash, falling/knockout motion, different fighting stance (aggressive, regressive and defensive) and customised weapon usage motion. Furthermore, 9 diverse non-violent activities between two NPCs are also present as control class. This includes: Yoga, gardening, dancing, exercising, conversation, argumentation, construction, vehicle-repairing, arrested.

USI dataset [21] on the other hand is primarily a social interaction dataset which contains 4 human interactions: **Talking**, **Shaking**, **Hugging** and **Fighting**. Three of these interactions are non violent, while the fighting category contains only cold hand-to-hand violence. It must be noted that no kind of hot fights is present in the USI dataset. However, similar to WVD this dataset contains interaction between a maximum of two people with a total of 16 videos per interaction of variable length. The USI dataset videos are also captured in a frontal or dash-cam view which is also similar to WVD. It must be noted that hugging and fighting classes are the most challenging to differentiate.

IV. HUMAN LABELS GENERATION: WVD AND USI DATASET

An important component of our methodology and essential part of our experimentation process involved the generation of human labels to evaluate the labels produced for the WVD and USI. To this end, for evaluation purposes, we have randomly sampled 10% of images from the WVD for human labelling as the test set. The validation set was 20% of the dataset. This result in 4,085 frames for test set, 8,169 frames for validation from a total of 40,845 images. The test set was then assigned to 15 participants for human labelling. Our general guideline was restricted to the detection of fights, for which each participant was asked to draw bounding boxes around the Region of Interest (RoI). All other aspects were left to human labellers to decide.

A few interesting observations have been made regarding the labelling behaviour of the participants. Each participant generated completely different bounding boxes for the same set. This resulted in different bounding box test set coordinates. This presents a challenge as to which test set should be considered and why. Further, what meaning would the performance reported on each test set hold? In order to solve this challenge and to include the variations of human labelling the test set was split into 6 (WVD) and 5 (USI) equal parts. A single test set was generated by combining labelled chunks from the labelled test sets. The reason for labelling variations is the inclusion of splattered blood, shadows and clothing accessories (such as hats or glasses) as part of the RoI, while others neglected these visual features completely. Furthermore, correctly labelling violent sequences in poor lighting conditions proved especially difficult for the human labellers, which increased inaccuracies in labelling. These

observations suggest that the understanding of a violent scene and bounding box coordinate generation are very subjective to humans. Figure 1 shows these variations in human labelling of the WVD; for the sake of clarity, only a sample of three bounding boxes is presented out of the total 6 for WVD. Each distinct color bounding box represents a human drawn bounding box on what they considered to be violence.

V. PROPOSED APPROACH

Violence detection is a very complex task since it requires the presence of certain visual and audio features such as blood, fire, weapons, screams or blasts. Additionally, motion and temporal features are widely utilised. In this work, our aim is to utilise simple (low level) object detection methods to generate bounding boxes for a (high level) violence detection task. We argue that high level understanding of events requires hierarchical structuring of low level information. Keeping this in mind, an automatic method is devised keeping human intervention in the labelling process to a minimum. The proposed method is divided into two stages.

- 1) **Stage 1:** In this stage, we utilise pre-trained object detection methods to generate weak labels by identification and aggregation of (Virtual NPCs) persons in the frames. Once aggregated the labelling quality is measured by calculating the IoU of the bounding boxes against human labelled WVD frames.
- 2) **Stage 2:** Using the weak labels generated through stage 1, state-of-the-art object detection methods are retrained to produce strong labels. the strong labels are evaluated against WVD [9] and USI human labelled test sets.

Figure 2 shows the two-stage process by which labels are generated. The figures show in depth the process by which weak labels are generated in stage 1. Afterwards, how these weak labels are utilised to produce a strong learner. The weak label generation can be seen in stage 1 of figure 2. 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 misclassification 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.

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

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

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.

A. Evaluation Criterion

Performance analysis of object detection methods can be divided into localisation and classification evaluation of the object. Mean Average Precision (mAP) [22] and IoU are the most commonly used evaluation matrices for these purposes. However, in this work, we are only interested in evaluating the localisation of violence. IoU is the *de facto* evaluation metric to measure the localisation performance of the model [23]. IoU, also known as Jaccard Index, simply takes the area of intersection and divides it by the area of union for the overlapping bounding boxes, as shown in Eq. 1. The IoU metric generates a normalised value between ‘0’ and ‘1’, where ‘0’ represents no overlap and ‘1’ represents complete overlap. IoU values are computed for images containing both human and automatic machine labels. This process is repeated for all the pre-trained object detection methods utilised in this work.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

B. Implementation Details

Since our approach uses pre-trained object detection methods, we utilise Tensorflow Object Detection API [24] and ImageAI [25] libraries to generate the machine labels for violence detection. Only the implementations with high detection rates were selected from the respective libraries. As the goal is to generate labels for violence, due of this only high performing models were selected. It must be noted that no new violence detection model is proposed in this work. However, through automatic labelling, our approach is able to produce high IoU values (i.e. localisation refer back to Section V-A), thus producing high detection scores.

The deep networks used in this work include **Tiny-Yolo** and **Yolo** [15], **Faster-RCNN (FRCNN)** [14], **RFCN** [16], **SSD** [17] and **RetinaNet** [18]. All these methods are pre-trained on the **MS-COCO** dataset [2], which is a large-scale dataset containing 80 different types of objects. The only object used in this approach is the “Person” class. These pre-trained networks are used to detect the persons in the WVD. As already mentioned, persons in the WVD are basically the NPCs of the photo-realistic game GTA-V. Therefore, despite these networks been trained to detect people using images of real world persons, they are utilised in our approach without any modification.

Once the person participating in the fights are detected, their bounding boxes are then used as anchors to generate larger bounding boxes for violence. For each successful detection, four values (x, y, w, h) are returned, where x and y are the coordinates of a point on the image, and w and h are width and height (respectively) for the bounding box with respect to the point (x, y) being the top-left corner of the bounding box.

To localise the RoI for the whole fight, the bounding boxes generated by the pre-trained object detection network

(for person detection) are first aggregated. The aggregation function utilised for it is given in Eq. 2. These bounding boxes are then saved and a record is maintained per violent scenario of the WVD. These records are utilised to replace bounding boxes in cases where the object detectors fails to detect the person. As fight sequences of the WVD contain violence captured with variant lighting conditions with NPC occlusion, deformation and scale variations (see Figure 1, 6, 7 and 8). Furthermore, as the object detectors are based on real world data, utilising them in a virtual setting can raise many false positives and miss classifications. To handle detection misses, previous bounding boxes are recorded and utilised by replacing a mean bounding box value. To this end, the previous three saved bounding boxes are utilised to calculate a mean bounding box value which is replaced when the object detector fails (cf Figure 2). Once the whole process is completed, the (semi accurate) weak labels and the IoU value for individual images and the dataset as a whole are stored. This marks the end of stage 1.

$$f_{agg} = (\min(x_1, x_2), \min(y_1, y_2), \max(x_1 + w_1, x_2 + w_2), \max(y_1 + h_1, y_2 + h_2)) \quad (2)$$

In stage 2, again using the Google object detection API, the models were retrained using the weak labels generated. In this stage, the learning rate of 0.0001 and early stopping strategy resulted in better IoU values. The early stopping forces the network to just learn the good labels as these labels are in majority which can be seen in Figure 3 while ignoring to fit bad labels during the training process, that resulted due to false positive in stage 1. Therefore, a low learning rate and early stopping strategy are instrumental in achieving high IoU scores at the completion of stage 2.

VI. RESULTS

Once the human labelling process was completed, and machine labels were generated for all the object detection methods mentioned in Section V-B. Figures 4 and 5 show the performance of each object detection method against each weapon type in the WVD. In the cold weapon category, object detection methods **FRCNN**, **SSD**, **Yolo**, **Tiny-Yolo** struggled for the “Golf Club and Knife” weapon categories. **RetinaNet** performed well overall. However, for hot weapons, “MGCombat” was the most challenging weapon category for **SSD** and **Tiny-Yolo**, while **FRCNN** only struggled on the “RifleCarbine” category. The performance for **RetinaNet** and **RFCN** looks slightly similar (Fig. 5). Since each scenario in the WVD has different scale variations (i.e. NPC occlusion and deformation), this causes unique IoU values for the same weapon, under different time settings.

To get a better understanding of the performance of object detectors, Figure 3 shows the frequency density of the labels produced for violence. We can clearly observe that both graphs, for hot and cold fights, are right skewed showing that most of the labels generated have a higher IoU value. However,

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

TABLE I
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

it must be emphasised that, in our method, hot labels have a greater chance of getting a higher IoU value. In Figure 3, **RetinaNet** and **RFCN** have shown their dominance under virtual settings for cold and hot fights, respectively.

Table I shows the overall performance of each object detection for the overall WVD. For hot fights, **RFCN** achieved an overall IoU value of **0.8256**, while **RetinaNet** achieved an IoU value of **0.7882**. Overall, for the whole dataset, **RetinaNet** had the highest IoU value of **0.8036**. This shows that our method has the ability to automatic labelling synthetic data such as the WVD dataset. This would enable the production of large-scale datasets for sensitive domains such as violence, further reducing human exposure during labelling and ethical implications.

A. The Good, The Bad, and The Ugly labels

After achieving high IoU values for our method of automatic labelling of violence, the labels were categorised into three classes: “The Good”, “The Bad” and “The Ugly”. The images with an IoU value higher than **0.8** are considered as “**The Good**”, the images with IoU ranging from **0.5 to 0.8** are considered as “**The Bad**”, and with IoU value less than **0.5** are considered as “**The Ugly**”.

Figure 6 shows “The Good” labels produced by our method against human labels (Section IV). The blue bounding boxes represent the latter, whereas the green bounding boxes represent the former. As evident in the figure, human and machine generated labels are near identical and completely overlapping. Furthermore, performance has been consistent against all the time settings of the WVD.

Figure 7 shows “The Bad” labels. Usually, in object detection, an IoU value above 0.5 is considered as *good*. However, in the case of violence detection, according to our view, IoU values of around 0.5 cannot be categorised as good. The

reason for this is due to the risk to life being higher in violence. The figure shows that human and machine labels are partially overlapping; here, a certain amount of violence can be detected, however, the machine labels are covering extra ground in most cases. In the case of “The Ugly” labels, shown in Figure 8, almost negligible or very little overlap is observed.

Upon closer observation of the images, poor performance can be attributed to three factors. First, the ratio of false positives is impacting the aggregation function (Eq. 2). The object detection methods were trained to detect people from real life data (MS-COCO), whereas we pitched them on a synthetic data captured in virtual settings without any pre-processing. This explains why certain objects were misclassified as *person* by the object detectors because of different learned feature representation. This can be observed in Figures 7 and 8, as certain extended machine generated bounding boxes show that some false positives were detected. Second, after successful violence bounding box generation, the coordinates are saved to predict the person in case of unsuccessful detection. The false positive violence bounding boxes cause irregularities. It can also be observed that, for the initial frames in the sequence, smaller records in the history also result in poor predictions. This indicates that an intelligent approach is required to predict the current violence bounding box rather than relying on previously recorded bounding boxes. Third, the NPC occlusion and deformation in consecutive frames cause the object detectors to completely miss one of the participants in the fights. Object detectors almost certainly fail on NPC that have fallen to the ground, or if they are out of the frame of capture or are subject to occlusion, as illustrated in Fig. 8. One such example of this trend can be observed for images of RetinaNet and YOLO, specifically for hot fights. Poor lighting conditions and NPC occlusion also affects labels of cold fights for Tiny-Yolo, SSD and RetinaNet. Here it can be clearly seen that a fallen NPC is often missed by the object

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.

detection methods. Further, occlusion with the background and deformation issues also affect performance. These factors affect the repository of previously recorded bounding boxes to predict missing person, thus is prone to error propagation. However, despite these challenges, our method produced good labels and the overall performance is also high.

B. Performance on USI Test Set

Once the training and evaluation were completed in stage 1. The good, bad and ugly labels were all randomly shuffled and split into training and validation sets and used to retrain the object detection methods to produce strong labels. The retraining process early stopping strategy was utilised so that

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

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

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

the method focuses more on the good labels rather than the bad or ugly ones. Once the retraining was complete the FRCNN gave the highest IoU score for the WVD validation set with a value of **0.95**. Once the strong labels and learners were generated. For final evaluation on a real world dataset. The USI human labelled test set was passed through the highest IoU producing method (i.e. FRCNN). The recorded performance is mentioned in Table II. The final IoU score achieved on the USI dataset is **0.7450**. “The Good”, “The Bad” and “The Ugly” labels for the USI datasets are shown in Figure 9. Here it can be seen that similar to WVD, labels generated cover most of the ROI for violence. Even the Bad labels cover most of the area where interactions between the individuals are taking place. This proves that the method is robust and can work for both virtual synthetic as well as for real world images with minimum human intervention. It must be noted that all the training was done on the virtual WVD dataset.

TABLE II
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

VII. FUTURE WORK

Our label generation approach applies to person-to-person fights; applying this technique for crowd fights would be an interesting avenue. Furthermore, the detection of person-to-person fights in a virtual crowd setting by training object detectors on the generated labels is another promising direction. As seen in our experimentation process that human labelling suffers from human psychological issues such as mind wandering or boredom. These human traits can add errors and variations generating unique bounding box coordinates. Deep labellers would solve this problem and able to generate a huge amount of labelled data in a limited time, with identical bounding box co-ordinates. Utilising low level features to build hierarchical representations to label high level tasks could prove useful for other computer vision tasks as well. However, such solutions could be custom build for specific domains, still automatic labelling to an acceptable error rate can still reduce human effort in data labelling.

The proposed method produced labels with high IoU values. However, in case of poor detection from the pre-trained networks, relying on the mean bounding box value calculated from previously generated bounding boxes is not a robust/smart approach. Moreover, errors generated and the quality of weak labels heavily impacts the performance. One possible solution is to use motion information of the fights; to predict the location of the NPCs. Furthermore, using recurrent networks has also proven to improve performance ([26]).

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

VIII. CONCLUSION

In this paper, we proposed a method for generating bounding boxes automatically for violence detection in a virtual environment by utilising object detection models pre-trained on the MS-COCO dataset. As part of our experimentation process 15 human participants labelled a subset of the Weapon Violence Dataset (WVD) and social interactions UNITN Social Interaction (USI) dataset. These test sets are utilised in our experimentation and will be publicly available.

Our approach for automatic labelling showed that extracting and combining low level information from simple object detectors can be used to label high level task of violence detection. Therefore, it indicated that Deep Learning-based object detection methods can be utilised as **deep labeller** to label a large set of images with superior IoU performance. The approach is useful to label sensitive images without or with minimum human involvement and exposure. Further, our experimentation showed Deep Learning methods trained on real world images can be effectively utilised on data captured in a virtual environment.

In stage 1, RetinaNet was able to produce an overall high quality weak labels with an IoU value of **0.8036** on the WVD while, specifically for hot and cold fights, RFCN and RetinaNet produced the highest IoU values of **0.8256** and **0.7882**. The overall trend for the bounding box generation

was positively skewed towards higher IoU values for all the DL methods used in our experimentation. In stage 2 this performance was further increased to **0.95** through passing the weak labels learned in stage 1. Finally, a single pass of the USI test set was performed, and an IoU score of **0.745** was recorded. This shows that our method can label images for violence detection with a high localisation accuracy with minimum human intervention.

ACKNOWLEDGMENT

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

REFERENCES

- [1] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

- [3] Rschmelzer, "Data preparation & labeling for ai 2020," Feb 2020. [Online]. Available: <https://www.cognilytica.com/2020/01/31/data-preparation-labeling-for-ai-2020/>
- [4] M. A. Killingsworth and D. T. Gilbert, "A wandering mind is an unhappy mind," *Science*, vol. 330, no. 6006, pp. 932–932, 2010.
- [5] J. D. Eastwood, A. Frischen, M. J. Fenske, and D. Smilek, "The unengaged mind: Defining boredom in terms of attention," *Perspectives on Psychological Science*, vol. 7, no. 5, pp. 482–495, 2012.
- [6] C. Demarty, B. Ionescu, Y. Jiang, V. L. Quang, M. Schedl, and C. Penet, "Benchmarking violent scenes detection in movies," in *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2014, pp. 1–6.
- [7] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR 2009-IEEE Conference on Computer Vision & Pattern Recognition*. IEEE Computer Society, 2009, pp. 2929–2936.
- [8] E. B. Nieves, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *International conference on Computer analysis of images and patterns*. Springer, 2011, pp. 332–339.
- [9] M. S. Nadeem, V. N. Franqueira, F. Kurugollu, and X. Zhai, "Wvd: A new synthetic dataset for video-based violence detection," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 2019, pp. 158–164.
- [10] J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang, and H. Wang, "Automatic player detection, labeling and tracking in broadcast soccer video," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 103–113, 2009.
- [11] T. Xiang and S. Gong, "Video behaviour profiling and abnormality detection without manual labelling," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1238–1245.
- [12] M. D. Bah, A. Hafiane, and R. Canals, "Deep learning with unsupervised data labeling for weed detection in line crops in uav images," *Remote sensing*, vol. 10, no. 11, p. 1690, 2018.
- [13] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari, "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*, 2016, pp. 854–863.
- [14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems* 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99. [Online]. Available: <http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf>
- [15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [16] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 379–387.
- [17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.
- [18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [19] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, no. 3, pp. 299–318, 2008.
- [20] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 8595–8598.
- [21] P. Rota, N. Conci, and N. Sebe, "Real time detection of social interactions in surveillance video," in *Computer vision—ECCV 2012. Workshops and demonstrations*. Springer, 2012, pp. 111–120.
- [22] L. LIU and M. T. ÖZSU, Eds., *Mean Average Precision*. Boston, MA: Springer US, 2009, pp. 1703–1703.
- [23] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [24] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7310–7311.
- [25] Moses and J. Olafenwa, "Imageai, an open source python library built to empower developers to build applications and systems with self-contained computer vision capabilities," mar 2018. [Online]. Available: <https://github.com/OlafenwaMoses/ImageAI>
- [26] Ş. Aktı, G. A. Tataroğlu, and H. K. Ekenel, "Vision-based fight detection from surveillance cameras," in *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2019, pp. 1–6.