HOD: New Harmful Object Detection Benchmarks for Robust Surveillance

Eungyeom Ha*  
Yonsei University, MODULABS

Heemook Kim*  
Inha University

Dongbin Na†  
POSTECH

Abstract

Recent multi-media data such as images and videos have been rapidly spread out on various online services such as social network services (SNS). With the explosive growth of online media services, the number of image content that may harm users is also growing exponentially. Therefore, the surveillance of these images is crucial. Thus, most recent online platforms such as Facebook and Instagram have adopted content filtering systems to prevent the prevalence of harmful content and reduce the possible risk of adverse effects on users. Unfortunately, computer vision research on detecting harmful content has not yet attracted attention enough. Users of each platform still manually click the report button to recognize patterns of harmful content they dislike when exposed to harmful content. However, the problem with manual reporting is that users are already exposed to harmful content. To address these issues, our research goal in this work is to develop automatic harmful object detection systems for online services. We present a new benchmark dataset for harmful object detection. Unlike most related studies focusing on a small subset of object categories, our dataset addresses various categories. Specifically, our proposed dataset contains more than 10,000 images across 6 categories that might be harmful, consisting of not only normal cases but also hard cases that are difficult to detect. Moreover, we have conducted extensive experiments to evaluate the effectiveness of our proposed dataset. We have utilized recent proposed state-of-the-art object detection architectures and shown our proposed dataset can be greatly useful for the real-time harmful object detection task. The codes and datasets are available at https://github.com/poori-nuna/HOD-Benchmark-Dataset.

1. Introduction

Recently, video and image content has been used in various online services. However, most online platforms and social media services are still monitoring the uploaded content in a post-processing manner. The problem with this post-processing approach is that users are exposed to harmful elements, which requires additional costs for organizations. Thus, recent websites have unprecedentedly needed detection systems for monitoring and regularizing harmful content. Intelligent cities monitor possibly dangerous objects such as knives and guns in real time [5, 16, 19]. However, most of the previously presented harmful object detection datasets have limitations in that they address a small subset of harmful object categories or provide only normal cases [15, 27, 28]. In this work, we present a new benchmark dataset for harmful object detection to overcome the limitations of previous studies. Our dataset contains more than 10,000 images over 6 categories. The dataset contains guns, knives, and diverse elements like alcohol, insulting gestures, blood, and cigarettes. The dataset includes normal cases and hard cases to detect. We further explore the effectiveness of our presented dataset and train the state-of-the-art architectures on our introduced dataset. We demonstrate the trained models achieve modest object detection performance. Our dataset provides individual hard case images, which is greatly useful for evaluating the robustness of harmful object detection algorithms. For the research purpose, we deploy all the presented datasets, source codes, and trained models. Our work provides the following main contributions.

• We present a novel harmful object detection dataset over 6 categories. To the best of our knowledge, our dataset covers the most various categories compared to the previous studies.

• Our dataset includes diverse hard cases that are hard to recognize and sometimes induce unexpected detection results, which are useful for evaluating the robustness of detection models. To evaluate the effectiveness of our proposed datasets, we have also extensively trained and evaluated the recent state-of-the-art object detection models on our datasets.

• We publicly provide all the datasets, source codes, and even the trained models for various online media services and surveillance systems to utilize our models as off-the-shelf methods easily.

*Correspondence to dongbinna@postech.ac.kr
†These authors contributed equally to this work.
2. Background and Related Work

2.1. Object Detection Using Deep Learning

Deep learning applications in computer vision have garnered significant attention due to their remarkable success across various industry domains [2, 16]. For image recognition tasks, the recently proposed deep-learning models based on convolutional neural networks (CNN) have shown improved classification performance, surpassing even humans [14, 21, 43, 47, 48]. These recently proposed deep CNN architectures suitable for extracting high-level semantic features can be used for various computer vision tasks such as semantic segmentation and object detection [29, 30, 37, 44]. Object detection methods based on R-CNN architectures have been used as baseline object detection models [17, 18, 39]. Faster R-CNN has relatively complex architectures. However, the detection performance is still competitive compared to the recently proposed methods [39, 49]. Faster R-CNN generally has been known to show competitive detection performance in that Faster R-CNN results in relatively lower false negatives (FN), although Faster R-CNN is relatively slower [38]. Therefore, a previous work uses Faster R-CNN as a baseline in FN-critical tasks such as gun and knife detection research [16]. YOLO-based methods have gained popularity for their simplicity and efficacy in real-time processing [38]. YOLO’s streamlined architectures show lower false positives (FP) during real-time detection, suitable for tasks demanding instant feedback [46].

2.2. Harmful Object Detection Dataset

In the general image object detection research fields, previous studies have presented various image datasets [13, 15, 28]. These datasets provide many image samples with bounding box annotations for various daily objects such as trucks, cars, etc. Despite its significance, the domain of harmful object detection datasets remains under-explored, specifically given the pressing demand in social media and online live-streaming services. In particular, online live-streaming service needs to reject harmful object that belongs to harmful categories, such as knives, blood, etc. Some previous studies have presented harmful object detection datasets [34]. However, most existing studies cover only a subset of the harmful categories or include only easy tasks. Moreover, the previous studies focus on other types of data, such as chemical signals or sensors, rather than images like ours [1, 4, 6, 25, 45]. The details of the studies using other datasets are described in the appendix. Therefore, we propose a new harmful object detection dataset covering 6 representative harmful categories and hard cases with extensive annotation effort of labeler participants.

3. Proposed Dataset

3.1. Category Selection and Annotation Criteria

Popular online platforms like Instagram, Twitter, and YouTube currently have strict content standards. However, their standards have tended to be somewhat focused on sexuality. Some previous studies argue that frequent exposure to violent elements can lead to aggression and desensitization to violence [23, 36, 42]. Therefore, as preventatives, blocking violent and potentially harmful elements that can lead to unexpected outcomes can be useful for various online platforms. In addition to violent objects, some studies have observed that visual elements that might be detrimental to users can potentially lead to negative consequences such as addiction and trauma [3, 8, 9, 12, 20, 35]. Thus, we have decided to select alcohol, insulting gesture, blood, cigarette, gun, and knife as a synthesis of prior research, social concerns, and the potential risks associated with exposure to these elements. A total of 5 labeler participants have collected a dataset of more than 10,000 images using search keywords based on the 6 categories: alcohol, blood, cigarette, gun, insulting gesture, and knife. A team of three main labelers has gathered over 1,500 images per category. As we have progressed through the experiments, another 2 labelers have collected additional images of underperforming categories. We note that each image can have two or more categories in a multi-label classification manner. The detailed labeling guide for each category is described in the appendix.

3.2. Data Distribution

We have divided the dataset into two distinct groups based on the difficulty of detection, the normal cases and the hard cases. The normal cases indicate easily identifiable images. These images are similar to datasets commonly utilized in existing research. However, our additional hard cases encompass images that are challenging to detect, which is a distinctive contribution from previous studies. Our hard case dataset mainly contains images of harmful objects that are small, or the objects’ category-discriminative features are covered. Hard cases also include images that the objects’ colors are similar to the background. Therefore, to infer the label of hard cases, we need other information, such as elements around the object or the context of the image. More details of the criteria for hard cases are demonstrated in the appendix. After splitting the entire dataset into the normal cases and hard cases, we split each dataset again into the training, validation, and test in a ratio of 8:1.5:0.5. Our extensive efforts ensure the absence of overlapping images between training, validation, and testing datasets with rigorous manual review processes. The number of data and examples per category for each case is described in Table 1 and Figure 1.
3.3. Rationale for Category Selection

- **Alcohol**: Numerous studies have highlighted the risks of early alcohol exposure, suggesting that exposure to alcohol objects can pave the way for substance misuse disorders in later life [20]. By identifying and moderating such content, we aim to mitigate the normalization of underage excessive alcohol consumption.

- **Blood**: Graphic visuals, particularly those displaying blood or gore, can sometimes induce fear and trauma in younger audiences [9]. Restricting access to such visuals may aid in fostering a safer media environment for children and adolescents.

- **Cigarette**: The World Health Organization (WHO) has persistently warned about the dangers of youth tobacco consumption, indicating that early exposure can lead to lifelong addiction [35]. By detecting and blurring such content, the allure and curiosity surrounding smoking might be reduced.

- **Gun**: Numerous studies have pointed out that exposure to firearms in media can influence aggressive behaviors and desensitize youth to real-life violence [8]. Therefore, by moderating this content, the objective diminishes the potential for gun-related curiosities and imitative behaviors.

- **Insulting Gesture**: Exposure to inappropriate or obscene gestures can mold the negative social behaviors of youngsters, often leading to the replication of such gestures in inappropriate situations [12]. The identification of these gestures aims to cultivate better behavioral norms.

- **Knife**: The representation of weapons, especially sharp ones like knives, has been correlated with an increased propensity for violent behaviors in young individuals [3]. Preventing young audiences from these visual triggers can potentially curtail the glamorization of violence.
3.4. Labeling Guidline

• Alcohol: Any bottle that could be recognized as a bottle of alcohol, no matter its shape, labelers have labeled the object as alcohol. We note that labelers have labeled not only the bottles but also any glasses that can be recognized as alcoholic beverages by the labelers.

• Blood: Labelers have labeled the objects as blood if blood is on the objects. Moreover, if blood is widely scattered, the labelers have grouped the blood regions and labeled them as a single blood object.

• Cigarette: Labelers have labeled all the individual cigarettes as possible. The labelers also have labeled e-cigarettes as cigarettes. Moreover, labelers have labeled cigarette packs as cigarettes.

• Gun: If labelers could recognize the object as a gun, whether the shape is a pistol, rifle, or sniper rifle, they have labeled the objects as guns.

• Insulting Gesture: The insulting gestures can appear differently according to the country. Therefore, after extensive discussion, the labelers agreed to define a specific finger shape commonly used worldwide as an insulting gesture. Specifically, the labelers have annotated the hand with only the middle finger extended and the rest of the fingers folded as an insulting gesture.

• Knife: Regardless of the type of knife, such as kitchen knives and long swords, labelers have labeled the objects as knives.

3.5. Hard Case Criteria

As we have noted, our hard cases contain images that are hard to detect. The criteria we have set are as follows. (1) Images that contain various objects with different categories in a multi-label manner. (2) Images that have small-sized harmful objects. The exact criteria for size are that both the width and height of the harmful objects relative to the image’s size are less than 0.2. (3) Images that have harmful objects whose category-discriminative features are unrecognizable. These cases include when the object’s color is similar to the background, the object is viewed from an unusual angle, or the category-discriminative features of the object are concealed. Additional criteria for each category are as follows.

• Alcohol: Images that are taken from above or below, rather than from the side, so that the shape of the bottle or glass can not be distinguishable.

• Blood: We note that the golden standard criteria of blood itself can be ambiguous. For example, scattered blood is grouped and labeled by the subjectivity of labelers. We have classified the images with multiple blood groups as hard cases.

• Cigarette: Images that the object has a relatively different shape, not a single cigarette, such as cigarette cases or electronic cigarettes.

• Gun: Images people can recognize as a gun, even though the barrel is heavily concealed. Additionally, we also consider the images of a gun that are viewed from the direction of the front barrel as a gun.

• Insulting gesture: Images taken from the side of the hand. Therefore, it is hard to tell if it is the middle finger.

• Knife: Images humans could recognize as a knife, even though the blades are not visible and only the handle is visible.

The representative images of hard cases our criteria apply are described in Figure 2. Moreover, we show the ground-truth image samples and the object detection results using the Faster R-CNN models in Figure 3. Similar to the YOLOv5 models, the Faster R-CNN models demonstrate improved object detection performance when using the hard case training dataset.

4. Experiments

We have utilized two baseline object detection architectures, YOLOv5 [37] and Faster R-CNN [39], representing one-stage and two-stage object detection methods, respectively. When reporting the main experimental results, we
Figure 3. The example images from our hard case test dataset and the corresponding inference results. The first row represents hard case ground-truth samples from the dataset $D_{test}^{hard}$. The second row shows the detection results using Faster R-CNN trained on only $D_{train}^{normal}$. The third row represents the detection results using Faster R-CNN trained on the joint dataset $D_{train}^{normal} \cup D_{train}^{hard}$. We have found hard case training images can greatly improve the detection models’ robustness.

Table 2. Detection performance of trained models. The table shows mAP scores per category.

<table>
<thead>
<tr>
<th>Train Dataset $D_{train}$</th>
<th>Test Dataset $D_{test}$</th>
<th>Models</th>
<th>Performance Measures</th>
<th>Categories</th>
<th>Alcohol</th>
<th>Insulting Gesture</th>
<th>Blood</th>
<th>Cigarette</th>
<th>Gun</th>
<th>Knife</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{train}^{normal}$</td>
<td>$D_{test}^{normal}$</td>
<td>YOLOv5</td>
<td>mAP@50 97.4, mAP@50-95 89.7</td>
<td>97.4, 89.7</td>
<td>97.8, 95.7</td>
<td>69.8, 48.6</td>
<td>79.9, 51.1</td>
<td>91.6, 80.5</td>
<td>95.0, 87.7</td>
<td>90.1, 76.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Faster R-CNN</td>
<td>mAP@50 89.3, mAP@50-95 72.3</td>
<td>73.6, 60.5</td>
<td>83.5, 62.5</td>
<td>90.0, 62.2</td>
<td>87.7, 62.2</td>
<td>97.2, 62.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_{train}^{hard}$</td>
<td>$D_{test}^{hard}$</td>
<td>YOLOv5</td>
<td>mAP@50 55.2, mAP@50-95 40.2</td>
<td>44.7, 26.2</td>
<td>41.2, 43.3</td>
<td>61.5, 39.4</td>
<td>47.9, 39.4</td>
<td>49.7, 37.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Faster R-CNN</td>
<td>mAP@50 39.0, mAP@50-95 24.5</td>
<td>20.1, 10.7</td>
<td>33.9, 22.0</td>
<td>38.0, 19.3</td>
<td>8.4, 19.3</td>
<td>19.8, 19.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_{train}^{normal} \cup D_{train}^{hard}$</td>
<td>$D_{test}^{normal}$</td>
<td>YOLOv5</td>
<td>mAP@50 99.2, mAP@50-95 92.8</td>
<td>97.1, 80.2</td>
<td>95.5, 86.9</td>
<td>98.4, 83.2</td>
<td>95.1, 83.2</td>
<td>94.5, 81.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Faster R-CNN</td>
<td>mAP@50 96.1, mAP@50-95 79.5</td>
<td>66.0, 44.5</td>
<td>85.0, 64.5</td>
<td>94.3, 63.7</td>
<td>89.7, 63.6</td>
<td>88.5, 63.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$D_{test}^{hard}$</td>
<td>YOLOv5</td>
<td>mAP@50 91.9, mAP@50-95 75.7</td>
<td>70.2, 46.3</td>
<td>88.2, 59.5</td>
<td>76.2, 55.4</td>
<td>74.9, 56.4</td>
<td>79.5, 59.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Faster R-CNN</td>
<td>mAP@50 83.1, mAP@50-95 57.9</td>
<td>57.4, 48.6</td>
<td>78.2, 53.3</td>
<td>64.9, 48.6</td>
<td>52.8, 48.6</td>
<td>66.9, 48.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

have adopted consistent hyperparameter settings for object detection models to obtain reliable and reproducible results.

4.1. Hyperparameter Tuning and Model Optimization

We have extensively experimented with various hyperparameters, such as batch size and weight initialization methods. For training YOLOv5 models, we use an image size of 416 and a batch size of 32. We have observed that the detection performance of the YOLOv5 models has converged after 200 epochs. For training Faster R-CNN models, we have adopted the MMDetection framework [11], notably known to provide baseline benchmarks in object detection research fields. We have experimented with various hyperparameter settings to leverage the detection framework fully. We have trained the Faster R-CNN models for 150 epochs with a learning rate of 0.0025, which demonstrates competitive detection performance in our harmful object detection tasks.

4.2. Model Training and Evaluation

We have trained object detection models and evaluated their effectiveness in various scenarios. Specifically, we have trained YOLOv5 and Faster R-CNN models using two distinct dataset configurations. The first setting contains...
only normal cases in the training dataset, while the second setting consists of normal and hard cases in the training dataset. Formally, in the first setting, we train the object detection models on only the normal case training dataset $D_{\text{train}}^{\text{normal}}$ and evaluate their detection performance on the normal case test dataset $D_{\text{test}}^{\text{normal}}$ and the hard case test dataset $D_{\text{test}}^{\text{hard}}$ individually. In the second setting, we train the object detection models on the joint training data distribution $D_{\text{train}}^{\text{normal}} \cup D_{\text{train}}^{\text{hard}}$ that consists of normal and hard cases during the training time. We expect that the second setting shows achieve improved generalization performance, which is desirable in that the object detection model trained on $D_{\text{train}}^{\text{normal}} \cup D_{\text{train}}^{\text{hard}}$ can capture more abundant feature representations of various difficult objects compared to models trained on the only normal case training dataset.

4.3. Performance Metrics

YOLOv5, which stands for “You Only Look Once,” is a one-stage, regression-based method for real-time object detection [37]. It offers end-to-end training, determining the target category and positioning simultaneously. The network structure consists of only convolutional layers and the input image layer. The YOLOv5 has been known for its lightweight and quick detection performance, surpassing other methods like Faster R-CNN in speed and precision benchmarks [7, 38]. The architecture of YOLOv5 is described in Figure 5.

We have evaluated our object detection models using mAP (Mean Average Precision). This mAP metric measures the comprehensive detection performance of an object detection model by calculating precision scores at different recall levels. Specifically, we have adopted 2 representative variations of the mAP: mAP@50-95 and mAP@95. We have comprehensively considered the various IoU thresholds and reported the overall detection results of trained models across 6 categories. The mAP scores can be individually calculated at each confidence threshold during evaluations. Therefore, we have thoughtfully selected the confidence thresholds that produce the highest mAP scores for each model utilizing the validation datasets. For YOLOv5 models, the optimal confidence score is 0.3, resulting in the highest mAP score.
for the test dataset. Meanwhile, for the Faster R-CNN models, the best confidence score is calculated as 0.1. We have meticulously assessed and compared the performance of our models [22, 28] by reviewing the confidence thresholds and mAP scores. We hope our experimental results provide valuable insights into the object detection research fields, significantly contributing to the advancement of harmful object detection research.

4.4. Model Architectures

We have utilized two baseline object detection architectures, YOLOv5 [37] and Faster R-CNN [39], representing one-stage and two-stage object detection methods, respectively. Faster R-CNN, with the Region Proposal Network (RPN), is a highly precise object detection method [39]. The approach comprises two stages, with the first stage utilizing a convolutional layer to extract features and generate feature maps. The region candidate network then creates candidate boxes, while the region of interest pooling layer collects feature maps and regional candidate frames. In the final stage, the classification layer identifies the object category and adjusts the position of the candidate frames. The VGG16 architecture can be utilized to detect small targets due to the low-resolution representations that are down-sampled and small pixel sizes on the feature space [41]. The structure of Faster R-CNN is described in Figure 6.

![Figure 6. The illustration of the Faster R-CNN network architecture.](image)

4.5. Analysis of Experimental Results

We note that the following four datasets do not overlap with each other. For training the models, we provide (1) normal case training dataset $D_{\text{train}}^{\text{normal}}$ and (2) hard case training dataset $D_{\text{train}}^{\text{hard}}$. For testing the models, we provide (3) normal case test dataset $D_{\text{test}}^{\text{normal}}$, (4) hard case test dataset $D_{\text{test}}^{\text{hard}}$. The main results of the experiments are presented in Table 2. We note that training the detection models on the joint dataset $D_{\text{train}}^{\text{normal}} \cup D_{\text{train}}^{\text{hard}}$ that contains normal and hard cases improves the overall detection performance across whole categories. The YOLOv5 models have improved the average mAP by 5, from 76.5 to 81.5, on the normal case test dataset $D_{\text{test}}^{\text{normal}}$. The category with the largest performance improvement is gun. The detection performance has been improved by 11.8, from 75.1 to 86.9. We note that the YOLOv5 model also shows significantly improved detection performance across whole categories by 22.2, from 37.4 to 59.6, on the hard case test dataset $D_{\text{test}}^{\text{hard}}$. The cigarette category shows the largest performance improvement. We have found that the detection performance has been improved by 36.9, from 26.2 to 63.1. We have observed that the Faster R-CNN model also achieves an improvement of the detection overall performance by 20.6, from 19.8 to 40.4, on the hard case test dataset $D_{\text{test}}^{\text{hard}}$ when using $D_{\text{train}}^{\text{normal}} \cup D_{\text{train}}^{\text{hard}}$. The largest performance gain occurs in the cigarette category, improved by 41, from 4.8 to 45.8. The examples of inference results on some hard case test samples are illustrated in Figure 4. Our experiments show hard case training dataset is crucial to achieving robust detection performance on various difficult objects.

5. Comparison with Existing Datasets

- **Alcohol**: Most existing alcohol detection work has been concerned with drunk driving [1, 6, 25]. These previous studies have not mainly aimed to detect alcohol images but to recognize the human body’s chemical signals or physical reactions after drinking alcohol. In the real world, alcohol can lead to unexpected accidents due to drunk driving. Therefore, we adopt the alcohol object category for our dataset. Likewise, avoiding frequent exposure to alcohol on the Internet can be considered important as a precaution.

- **Blood**: Existing studies on blood detection are mainly based on medical or forensic perspectives [4, 45]. The previous datasets are frequently used to detect diseases through blood tests or to detect bloodstains on crime scenes and evidence using the luminol chemical reaction. Therefore, the existing work either has dealt with blood images on a cellular level or uses data from chemical sensors. The bloodstain images themselves have not been frequently treated as datasets.

- **Cigarette**: The purpose of traditional cigarette detection studies is to prevent risky incidents caused by smoking behavior. Their data samples generally include fire signals caused by smoking at gas stations and accidents caused by smoking while driving [10, 26, 31]. Therefore, many studies have focused on detecting cigarette smoke. Different from previous studies, we aim to detect the cigarette object itself.

- **Sign language**: The existing finger pose research addresses sign language detection for hearing-impaired people [32, 40]. Their work has focused on detecting and interpreting what the sign language represents based on estimating the pose of a person’s fingers. They generally utilize tilt and accelerometer sensors on the
fingers through data gloves. However, most existing studies do not deal with images of insulting hand gestures used by hearing-abled people. We are the first to address and provide a dataset of insulting hand gestures in images.

- **Weapon**: Existing studies on harmful elements mainly focus on weapons such as guns and knives [24,33]. Unlike our study, previous studies generally have intended to prevent real-world violence and terrorism, not to detect harmful elements in internet content. Gun and knife detection is necessary to prevent terrorism in the real world. Therefore, we also adopt these categories, yet focus on the Internet media content to collect images. We note that many studies claim that frequent exposure to weapons such as guns and knives leads to familiarity with them [3,8]. Thus, we also consider these categories as harmful on the Internet as a precaution.

6. **Discussion**

Our goal is to develop an automated classification system to detect potentially harmful objects and prevent exposure to the harmful objects. Therefore, the detection model should be able to detect harmful objects in various cases, including normal and hard cases. The final goal is to train many hard cases so that models can detect harmful elements like humans, even when the distinguishing features of harmful objects are concealed. The ideal object detection model can detect harmful objects based on the overall context of the images, even if the most significant features of the harmful elements are masked. Figure 8 is an example of a knife in which the blade of the knife is hidden. This example is a representative scene that needs to be detected as a knife. However, when training object detection models only on normal case images, the hard case objects can not be easily detected. Thus, we will continue collecting, training, and deploying additional hard cases to detect harmful objects, even in hard cases like Figure 8 for future work.

6.1. **Sources of Images and Labeling Tools**

Our team of labelers has crawled images from the following websites where users can use photos for research purposes. Labeler participants have utilized One Click Image Downloader for crawling and Make Sense for labeling. Labelers set the whole categories first, then annotate and label all images. Labelers have utilized rectangular bounding boxes for the annotation. After the labelers finished labeling, they exported annotation files in YOLO and VOC formats. The actual procedure conducted by the labelers is illustrated in Figure 7.

7. **Conclusion**

This paper presents a new benchmark valuable dataset for the harmful object detection task, helping various users and organizations automatically address potentially harmful visual content elements. We provide datasets that cover 6 categories: alcohol, blood, cigarette, gun, insulting gesture, and knife. For constructing the harmful object detection dataset, we have first chosen normal cases of images similar to the datasets used in previous studies. However, we note that the hard cases that are hardly recognizable frequently induce unexpected model outputs. With extensive experiments, we have demonstrated training the detection model on the normal and hard cases simultaneously shows improved detection performance over all 6 categories. The experimental results conclude that hard cases training samples are beneficial for recognizing various shapes of harmful objects. We hope our presented datasets, trained models, and source codes can be utilized for various online services and research fields that adopt visual censorship systems.

8. **Acknowledgements**

This research was supported by Brian Impact, a non-profit organization dedicated to advancing science and technology.
References


[27] Sakib B Kibria and Mohammad S Hasan. An analysis of feature extraction and classification algorithms for dangerous


