Object Detection Binary Classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance

Article *in* Knowledge-Based Systems · February 2020

DOI: 10.1016/j.knosys.2020.105590

Francisco Pérez
University of Granada

Siham Tabik
University of Granada

Alberto Castillo Lamas
University of Granada

Object Detection Binary Classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance

Francisco Pérez-Hernández[a,*], Siham Tabik[a], Alberto Lamas[a], Roberto Olmos[a], Hamido Fujita[b], Francisco Herrera[a,c]

[a] *Andalusian Research Institute in Data Science and Computational Intelligence, University of Granada, 18071 Granada, Spain*
[b] *Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh City, Vietnam*
[c] *Faculty of Computing and Information Technology, King Abdulaziz University (KAU) Jeddah, Saudi Arabia*

Abstract

The capability of distinguishing between small objects when manipulated with hand is essential in many fields, especially in video surveillance. To date, the recognition of such objects in images using Convolutional Neural Networks (CNNs) remains a challenge. In this paper, we propose improving robustness, accuracy and reliability of the detection of small objects handled similarly using binarization techniques. We propose improving their detection in videos using a two level methodology based on deep learning, called Object Detection with Binary Classifiers. The first level selects the candidate regions from the input frame and the second level applies a binarization technique based on a CNN-classifier with One-Versus-All or One-Versus-One. In particular, we focus on the video surveillance problem of detecting weapons and objects that can be confused with a handgun or a knife when manipulated with hand. We create a database considering six objects: pistol, knife, smartphone, bill, purse and card. The experimental study shows that the proposed methodology reduces the number of false positives with respect to the baseline multi-class detection

*Corresponding author

Email addresses: fperezhernandez@ugr.es (Francisco Pérez-Hernández), siham@ugr.es (Siham Tabik), albertocl@ugr.es (Alberto Lamas), h.fujita@hutech.edu.vn (Hamido Fujita), herrera@decsai.ugr.es (Francisco Herrera)

model.

Keywords:
Detection, Convolutional Neuronal Networks, One-Versus-All, One-Versus-One

1. Introduction

Many real world problems require the detection of multiple objects in images or videos [30]. Building useful detectors for such problems can be solved using modern deep learning models especially when the target objects are different, i.e., different size, colour, shape and texture. However, this task becomes more complicated when the target objects are small (represented by a reduced number of pixels, similar size, shape, colour and texture) and handled similarly.

Currently, the most accurate detection models are based on deep Convolutional Neural Networks [22, 42]. These models automatically learn the distinctive features of objects from a large set of labelled data [40]. The detection model that won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [35] in 2017 achieved a mean average precision of around 73% on a dataset of 527,892 images organised into 200 object classes [17]. The detection model that won the Common Objects in Context (COCO) challenge in 2017 achieved an average precision of around 73% on a dataset organised into 80 classes. The largest average precision, 66%, and largest average recall, 82%, were achieved on large objects while the lowest average precision, 34%, and average recall, 52%, was obtained on small objects[1] [24]. In general, robust detection models combine a meta-architecture, such as Faster-RCNN or R-FCN [33, 8], with one of the state-of-the-art classification architectures based on ResNet, VGG or Inception [16, 36, 37].

The capability to distinguishing between several small objects manipulated with hand is essential in several fields, especially in video surveillance, where the correct detection is extremely important. An important case study for violence

[1] http://cocodataset.org/#detection-leaderboard

prevention is the detection of weapons in places such as, banks or jewelleries, where people often handle objects that can be confused with a handgun or a knife as they are handled similarly, smartphone, bill, purse and card.

On the other hand, binarization techniques such as One-Versus-All (OVA) [6, 3] and One-Versus-One (OVO) [21, 32, 2] convert a multi-class problem into several expert binary models and calculate the final class using an aggregation method. These techniques are often used to reduce the instability in imbalanced problems [41, 11] and they present a good potential for the problem of similar objects detection.

This work proposes an accurate and robust methodology, Object Detection with Binary Classifiers based on deep learning (ODeBiC methodology), for the detection of small objects manipulated similarly with hand applied to surveillance videos.

The first model for weapon detection in videos was proposed by Olmos et al. [27]. The authors formulated the problem into a two-class (pistol and background) problem, built a training database using images from Internet and used Faster-RCNN based on VGG16 [27] as detection model. In general, this model reaches good results, but confuses the pistol with objects that can be handled similarly, for example, knife, smartphone, bill, purse and card. Figure 1 shows some of these false positives. This results show that the way in which pistols are handled is considered by the model as key feature of the pistol class, which is a problem from the video surveillance point of view. We address this case study with the ODeBiC methodology, with the aim of improving the detection among small objects handled similarly.

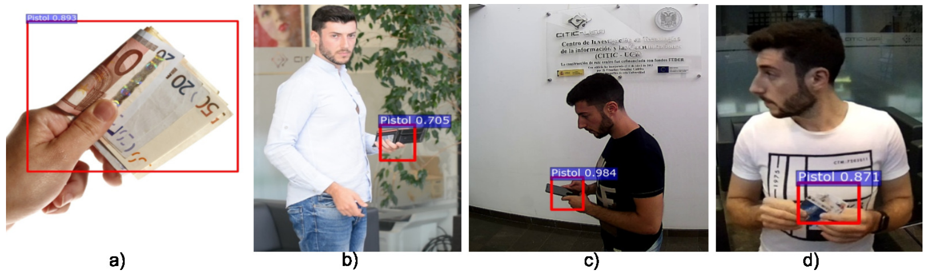

Figure 1: False positives committed by the proposed model in [27], where the objects are a) bill, b) purse, c) smartphone and d) card.

The main contributions of this paper are:

- We propose and evaluate a two level methodology called ODeBiC, based on the use of deep learning, to improve the detection of small objects that can be handled similarly. The first level uses a detector to select from each input frame the candidate regions with a specific confidence about the presence of each object. Then, the second level analyses these proposals using a binarization technique to identify the objects with higher accuracy. ODeBiC methodology maintains a good accuracy for the detection of large objects as well.

- We analyse the potential of binarization techniques such as, OVA and OVO, to improve the detection of small objects, manipulated with hand, that can be confused with a weapon. As far as we know, this is the first study in analysing such potential.

- We build a new dataset called Sohas_weapon (small objects handled similarly to a weapon, dataset) for the case study of six small objects that are often handled in a similar way to a weapon: pistol, knife, smartphone, bill, purse and card. We used different camera and surveillance camera technologies to take the images. 10% of the images were downloaded from Internet. All these images were manually annotated for the detection task.

This useful dataset will be available for other studies[2].

Our experimental study on the database Sohas_weapon applying the ODeBiC methodology overcomes the baseline detection model by up to 19,57% in precision and reduces the number of false positives by up to 56,50%.

This paper is organised as follows. Section 2 includes related works and preliminaries of the binarization techniques and object detection. Section 3 provides a description of the database construction and the test surveillance videos used to analyse the methodology and the proposed ODeBiC methodology. Section 4 gives the experimental analysis and comparison of ODeBiC methodology with different classification approaches. Finally, conclusions and future works are given in Section 5.

2. Binarization techniques and object detection

This section is organised into two parts. Subsection 2.1 provides a summary of related works that use binarization strategies, the state-of-the-art in object detection in images and the studies that address weapon detection in videos. Then, it presents a brief summary of OVA and OVO binarization methods in subsection 2.2 and 2.3 respectively.

2.1. Related works on binarization for objects detection in images

Related works can be divided into three categories: previous works that use OVA and OVO binarization strategies in classification, detection or segmentation, the state-of-the-art of object detection models in images and previous works that address weapon detection in videos.

Most prior works that analysed OVA and OVO in visual tasks, object recognition, image classification and image segmentation, only use classical models such as Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and k-Nearest Neighbours (kNN):

[2]http://sci2s.ugr.es/weapons-detection

- In image classification, the authors in [34] analysed OVA and OVO approach to reduce the features space on three well known benchmarks, MNIST, Amsterdam Library of Object Images (ALOI) and Australian Sign Language (Auslan).

- For pose estimation using image segmentation, the authors in [39] compared an individual CNN-based classifier with OVA and OVO based on SVM and showed that CNNs achieves slightly better performance than OVA and OVO based on SVM.

- Similarly, in the task of scene classification in remote sensing images, the authors in [5] also compared OVA and OVO based on SVM and 1-NN (Nearest Neighbour) and concluded that OVA provided worse results due to the unbalance between classes. The best results were obtained by OVO based on SVM.

- In face recognition, the authors in [29] used a CNN-based model for features extraction and an SVM, OVA and OVO for classification. The best results were obtained by CNN combined with SVM.

- The authors in [23] compared the Half-Against-Half (HAH) technique with OVA and OVO in image classification and found that HAH provides similar or worse results on the evaluated benchmarks.

On the other hand, we must highlight that the state-of-the-art detection models are end-to-end CNNs that combine a detection meta-architecture with a classification model. The most influential meta-architectures are Faster-RCNN [33], R-FCN [8] and SDD [25]. According to [18], Faster-RCNN based on Inception ResNet V2 obtains the highest accuracy on large objects while Faster-RCNN ResNet 101 provides the highest accuracy on small objects. SSD is the fastest detection approach but offers lower accuracies. The model that provide the best trade-off accuracy and execution time is Faster-RCNN ResNet 101.

In video surveillance, the first pistol detection model was proposed in [27], it provides good results but produces an important number of false positives in the background class due to the fact that the model confuses the pistol with objects that are handled similarly to a pistol. The authors in [26] propose a fusion technique with the support of two symmetric cameras to calculate the disparity map then subtract the background and consequently decrease the number of false negatives in the background. In the same direction, the authors in [4] reduce the number of false negatives produced by the extreme light conditions using a brightness guided pre-processing method.

Our current work is different from all the previously cited works in that it aims at developing a methodology that reduces the number of false positives and improves the overall performance in the detection of small objects handled similarly. As case study, we address the problem of identifying small objects handled similarly to a weapon in surveillance videos. As far as we know, this work is the first in applying OVA and OVO to deep learning models for object detection in images and videos.

2.2. One-Versus-All (OVA)

OVA strategy [6, 3] reformulates the multi-class classification problem into a set of binary classifiers where each classifier learns how to distinguish each individual class versus all the rest of classes together. This approach produces as many classifiers as the number of classes in the original problem. The final prediction is calculated by combining the predictions of individual classifiers using an aggregation method called Maximum confidence strategy (MAX). The class with the largest vote is considered as the predicted class. Formally, the MAX decision rule can be expressed as,

$$Class = arg \max_{i=1,\ldots,m} r_i, \tag{1}$$

where $r_i \epsilon [0, 1]$ is the confidence for class i and m is the number of classes.

2.3. One-Versus-One (OVO)

OVO strategy [21, 32, 2] translates the original multi-class problem into as many binary problems as all the possible combinations between pairs of classes so that each classifier learns to discriminate between each pair. That is, a $m-$class problem will be converted into $m(m-1)/2$ classifiers. In the specific case considered in this work with $m = 6$ will be translated into 15 classifiers. The prediction produced by all the classifiers will be combined in a confusion matrix and analysed by an aggregation method.

OVO system can use diverse aggregation strategies. Namely, the Max-Wins rule (VOTE), Weighted voting strategy (WV), Learning valued preference for classification (LVPC), Preference relations solved by Non-Dominance Criterion (ND), Classification by pairwise coupling (PC), Wu, Lin and Weng probability estimates by pairwise coupling approach (PE) and Distance-based relative competence weighting combination for OVO (DRCW-OVO).

VOTE rule

VOTE rule, also called Max-Wins rule [12], is considered as the basic decision rule in OVO. It analyses each element r_{ij} of the confusion matrix, if the prediction r_{ij} is equal or larger than 0.5, the output class will be i, on the contrary the output class will be j. The result is summed and the class with the larger votes is selected. If we have two or more classes with the same number of votes, we propose two alternatives:

- VOTE random: select one randomly.
- VOTE by weight: sum the predictions and select the maximum class value as the final class.

Formally, the decision rule can be written as:

$$Class = arg \max_{i=1,\ldots,m} \sum_{i \leq j \neq i \leq m} s_{ij}, \tag{2}$$

where s_{ij} is 1 if $r_{ij} > r_{ji}$ and 0 otherwise.

Weighted voting strategy (WV)

The aim of this technique [13] is to obtain the class with the largest probability. Hence, each class sums it predictions and the class with the maximum value is the final result. The decision rule is:

$$Class = arg \max_{i=1,\ldots,m} \sum_{i \leq j \neq i \leq m} r_{ij} \tag{3}$$

Learning valued preference for classification (LVPC)

Learning valued preference for classification (LVPC) technique calculates some new values from the initial probabilities obtained by the binary classifiers. LVPC is a weighted voting, it penalises the classifiers that have not got a threshold confidence in their decision. More details on this rule are provided in [20, 19]. This decision rule can be expressed as:

$$\begin{aligned}
P_{ij} &= r_{ij} - min\{r_{ij}, r_{ji}\} \\
P_{ji} &= r_{ji} - min\{r_{ij}, r_{ji}\} \\
C_{ij} &= min\{r_{ij}, r_{ji}\} \\
I_{ij} &= 1 - max\{r_{ij}, r_{ji}\} \\
Class &= arg \max_{i=1,\ldots,m} \sum_{i \leq j \neq i \leq m} P_{ij} + \frac{1}{2} C_{ij} + \frac{N_i}{N_i + N_j} I_{ij},
\end{aligned} \tag{4}$$

where N_i is the number of examples from class i in the training data.

Preference relations solved by Non-Dominance Criterion (ND)

The ND technique, also called, Preference relations solved by Non-Dominance Criterion, was initially introduced in decision making with fuzzy preference relations [28, 10]. The same criterion can be applied to an OVO classification system.

First, we should normalise:

$$\overline{r}_{ij} = \frac{r_{ij}}{r_{ij} + r_{ji}} \tag{5}$$

Then, compute the fuzzy strict preference:

$$r'_{ij} = \begin{cases} \overline{r}_{ij} - \overline{r}_{ji}, & when\ \overline{r}_{ij} > \overline{r}_{ji} \\ 0, & otherwise. \end{cases} \tag{6}$$

And, compute the non-dominance degree of each class:

$$ND_i = 1 - \underset{j \epsilon C}{sup}[r'_{ji}] \tag{7}$$

Finally, the output:

$$PredictedClass = arg \max_{i=1,...,m} ND_i \tag{8}$$

Classification by pairwise coupling (PC)

The PC technique or Classification by Pairwise coupling [15] attempts to enhance the voting strategy when the outputs of the classifiers are probabilities. This method calculates the joint probability for all classes from the pairwise class probabilities of the binary classifiers.

The proposed algorithm was:

1. Initialisation:

$$\begin{aligned} \widehat{p}_i &= \frac{2}{m} \frac{\sum\limits_{1 \le j \ne i \le m} r_{ij}}{(m-1)} for\ all\ i = 1, ..., m \\ \widehat{\mu}_{ij} &= \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{p}_j} for\ all\ i, j = 1, ..., m \end{aligned} \tag{9}$$

2. Repeat until convergence:

    (a) Compute $\widehat{p}$

$$\widehat{p}_i = \widehat{p}_i \frac{\sum\limits_{1 \le j \ne i \le m} n_{ij} r_{ij}}{\sum\limits_{1 \le j \ne i \le m} n_{ij} \widehat{\mu}_{ij}} for\ all\ i = 1, ..., m \tag{10}$$

where n_{ij} is the number of training data in the ith and jth classes.

(b) Normalise $\widehat{p}$

$$\widehat{p}_i = \frac{\widehat{p}_i}{\sum\limits_{i=1} \widehat{p}_i} for\ all\ i = 1, ..., m \tag{11}$$

(c) Recompute $\widehat{\mu}_{ij}$

$$\widehat{\mu}_{ij} = \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{p}_j} for\ all\ i, j = 1, ..., m \tag{12}$$

Finally, the output class:

$$Class = arg \max_{i=1,...,m} \widehat{p}_i \tag{13}$$

Wu, Lin and Weng probability estimates by pairwise coupling approach (PE)

The PE technique, also called Wu, Lin and Weng probability, is similar to PC. It uses the pairwise coupling approach to calculate the predictions [38]. The probabilities (p) of each class are estimated starting from the pairwise probabilities. PE optimises the following problem:

$$\underset{min}{p} \sum_{i=1}^{m} \sum_{1 \leq j \neq i \leq m} (r_{ji} p_i - r_{ij} p_j)^2\ subject\ to \sum_{i=1}^{k} p_i = 1, p_i \geq 0, \forall i \tag{14}$$

Distance-based relative competence weighting combination for One-Versus-One (DRCW-OVO)

Distance-based relative competence weighting combination, also called One-Versus-One strategy in multi-class problems (DRCW-OVO) [14], is one of variations [7] of OVO technique that intends to improve the problem of the imbalanced classes using the distance with the k elements near of the new instance.

Once the score-matrix has been obtained, DRCW-OVO entails the following:

1. Calculate the average distance of the k nearest neighbours of each class in a vector **d**.

2. Calculate the new score-matrix R^w as follows:

$$r_{ij}^w = r_{ij} \cdot w_{ij}, \tag{15}$$

where w_{ij} is computed as:

$$w_{ij} = \frac{d_j^2}{d_i^2 + d_j^2}, \tag{16}$$

being d_i the distance of the instance to the nearest neighbour of class i.

3. Use Weighted voting strategy (WV) on the new score-matrix R^w to obtain the final class.

Our problem is that we work with images, and we need calculate the distance. For this reason, a form to do it, could be calculate the Quadratic-Chi [31] with the histogram of the images:

$$X^2(P,Q) = \frac{1}{2} \sum_i \frac{(P_i - Q_i)^2}{(P_i + Q_i)}, \tag{17}$$

where P_i is the histogram of the new instance and Q_i is the average of the histogram of the k nearest neighbours.

3. Sohas_weapon database and ODeBiC methodology based on deep learning

We propose the ODeBiC methodology based on deep learning for binary classifiers with the aim to detect small objects that can be confused because they are handled similarly. As case study, we select a problem from the field of video surveillance, the detection of small objects that can be confused with a pistol or knife. We create the datasets called Sohas_weapon.

In this section, first we describe the process we used to build a dataset of small objects that can be hold similarly (subsection 3.1). Then, we present the ODeBiC methodology (subsection 3.2).

3.1. Sohas_weapon database construction for detection in surveillance videos

The quality of the learning of a CNN model depends strongly on the quality of the training database. The database must allow the classification model to correctly distinguish between objects handled similarly.

We built four databases for training the classifications models, Database-1, 2, 3 and 4 using different types of images. These databases are based on the case study of the similar handled objects like pistol, knife, smartphone, bill, purse and card:

1. In the first step, we used the pistol images from the database[3] built in [27] and the knife images from the database built in [4]. Most images were downloaded from Internet. We added the images of common objects that can be handled similarly to a pistol and a knife: smartphone, bill, purse and card. This database will be called Database-1.
2. In a second step, we added to each class images taken in diverse conditions by a reflex camera, Nikon D5200. The obtained database will be called Database-2.
3. In a third step, we added to each object class images taken by two surveillance cameras with different qualities and resolutions, Hikvision DS-2CD2420F-IW and Samsung SNH-V6410PN, and under diverse conditions. The obtained database will be called Database-3.
4. In the last step, we eliminated blurry images due to the motion and images where the human eye cannot recognise the object class. As we have mentioned the final database will be called Sohas_weapon.

To evaluate the quality of the databases guided by the quality of the learning of the classification approaches we built a database called Database-Sohas_weapon-Test. The characteristics of all the built databases are provided in Table 1. Besides, we used a database without pistol and knife class, Database-Sohas_weapon-Without_Pistol&Knife and Database-Sohas_weapon-Test_Without_Pistol&Knife,

[3]http://sci2s.ugr.es/weapons-detection

to analyse the behaviour of the proposed classification approaches on the objects that have a higher similarity in shape and way in which they are handled, smartphone, bill, purse and card.

To training the detection models, we used Database-Sohas_weapon-Detection whose characteristics are summarised in Table 1. This database contains the entire images (objects and background) from which we cropped the images used to build the database Sohas_weapon.

Table 1: Databases built to analyse the performance of objects that are manipulated similarly with hand.

Database-	**# img**	**Pistol**	**Knife**	**Smartphone**	**Bill**	**Purse**	**Card**
1	6589	3394	1879	866	134	137	179
2	7333	3523	1879	1022	287	315	307
3	8537	3681	1879	1069	654	710	544
Sohas_weapon	5680	1580	1879	755	545	581	340
Sohas_weapon-Without_Pistol&Knife	2221	0	0	755	545	581	340
Sohas_weapon-Detection	5080	1425	1825	575	425	530	300
Sohas_weapon-Test	1170	294	470	115	123	104	64
Sohas_weapon-Test_Without_Pistol&Knife	406	0	0	115	123	104	64

To analyse ODeBiC methodology, we created four test surveillance videos whose characteristics are summarised in Table 2. These four surveillance videos were recorded in different scenarios: in a small office and in a hall at the entrance of a building, in their viewpoints of the hall, with Samsung SNH-V6410PN camera.

Table 2: Four test surveillance videos created to analyse the performance of ODeBiC methodology.

Video	**# Frames**	**Pistol**	**Knife**	**Smartphone**	**Bill**	**Purse**	**Card**	**Scenario**
1	1962	235	289	217	302	342	391	Small office
2	2083	269	256	477	282	294	417	Hall view left far
3	2070	329	274	284	294	330	356	Hall view left near
4	2188	315	246	458	323	331	504	Hall wall

3.2. ODeBiC methodology based on deep learning

One of the main issues in object detection in surveillance videos is that the objects that can be handled similarly can be confused. This was shown in the pistol against background detection model developed in our previous work [27].

Herein, we propose using ODeBiC methodology based on deep learning to improve the reliability, robustness and accuracy to identify small objects handled similarly. ODeBiC methodology has two level, the first level obtains candidate regions that contain the target objects, and the second level classifies each region with the binarization technique followed by an aggregation method to finally produce the output frame with the detection results. In particular, ODeBiC methodology works as follows:

- The first level analyses the input frame using a relaxed CNN-detection model that outputs all the region proposals with a probability of having one or more target objects higher than 10%. This process could be seen as a candidate selection technique with an important knowledge of the target object categories. We will consider Faster-RCNN based on ResNet101 feature extractor as it provides a good trade-off between accuracy and execution time. These candidates will be analysed by the second level.

- Each output box will be analysed by a binarization technique, then an aggregation method is applied to calculate the final prediction. We will consider two binarization techniques, OVA and OVO, in combination with different aggregation methods. An illustration of OVA and OVO in the context of the pistol or knife and similar objects problem is depicted in Figure 2 and Figure 3 respectively.

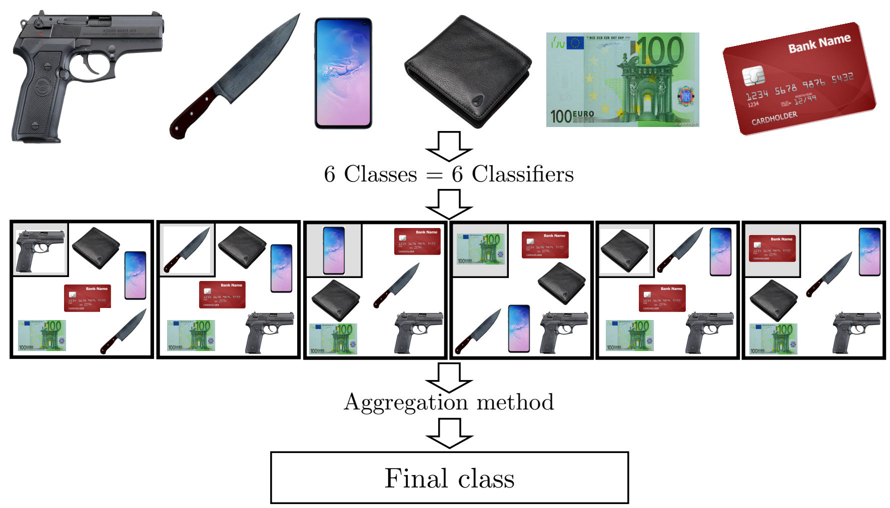

Figure 2: OVA process in the problem of recognising small objects that can be manipulated with hand in a similar way.

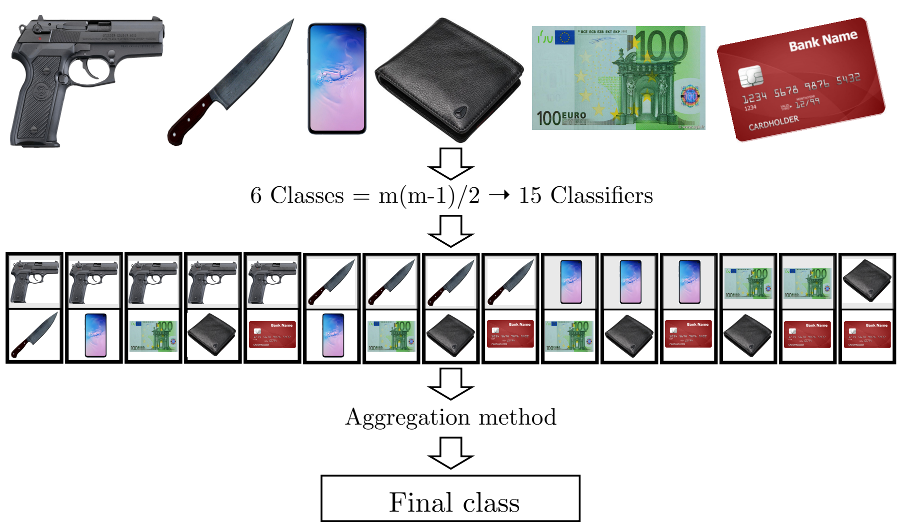

Figure 3: OVO process in the problem of recognising small objects that can be manipulated with hand in a similar way.

The proposed two level methodology is depicted in Figure 4.

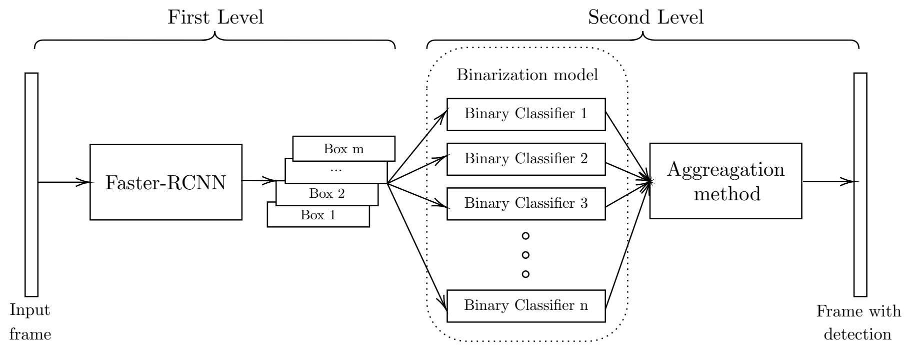

Figure 4: The structure of the proposed two level methodology, ODeBiC, first detection then binarization.

4. Experimental study

The purpose of this section is to analyse the performance of different classification approaches, the baseline multi-classifier, OVA and OVO with several aggregation rules in subsection 4.1, the study of similar objects in subsection 4.2 and the evaluation of our methodology ODeBiC using four surveillance videos in subsection 4.3.

4.1. Evaluation of different classification approaches

In this subsection we analyse the performance of different classification approaches, the baseline multi-classifier, OVA and OVO with different aggregation rules, VOTE random, VOTE by weight, WV, LVPC, ND, PC, PE, DRCW with k=1, 2, 3 and 4, trained on Databases-1, 2, 3 and Sohas_weapon and tested on Sohas_weapon-Test. All the analysed CNN models are based on ResNet-101 architecture [16] initialised with the pre-trained weights on ImageNet [9]. We used TensorFlow [1] and NVIDIA Titan Xp for all the experiments. The training process takes approximately two hours. The results are plotted in Figure 5 and summarised in Table 3.

Table 3: Results of all the classification approaches trained on Database-1, 2, 3 and Sohas_weapon and tested on Database-Sohas_weapon-Test.

	Database-1			**Database-2**			**Database-3**			**Database-Sohas_weapon**			**Time (s)**
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	
Baseline multi-classifier	86.38%	77.35%	79.88%	86.87%	82.67%	84.18%	90.02%	89.42%	89.62%	91.30%	91.03%	91.09%	0,02821
OVA	87.02%	75.56%	78.71%	88.29%	82.40%	84.50%	90.49%	89.15%	89.72%	92.76%	92.03%	92.29%	0,03081
OVO VOTE random	85.29%	74.94%	78.23%	83.79%	78.87%	80.30%	91.59%	91.32%	91.38%	93.68%	**93.16%**	93.35%	0,02824
OVO VOTE weight	86.18%	75.40%	78.61%	85.35%	79.94%	81.67%	92.00%	91.54%	91.70%	93.85%	92.96%	93.35%	0,02823
OVO WV	85.95%	75.44%	78.60%	85.69%	80.23%	81.97%	91.44%	91.27%	91.29%	93.45%	92.68%	93.01%	0,02822
OVO LVPC	86.20%	74.15%	77.69%	85.35%	79.24%	81.28%	92.25%	91.32%	91.70%	93.55%	92.55%	93.00%	0,02828
OVO ND	85.50%	74.67%	77.81%	85.24%	80.25%	81.86%	91.86%	91.38%	91.55%	**93.87%**	93.09%	**93.43%**	0,02827
OVO PC	86.15%	73.98%	77.12%	84.70%	80.00%	81.34%	91.25%	90.97%	91.04%	93.41%	92.84%	93.07%	0,04493
OVO PE	84.84%	74.27%	77.37%	85.09%	79.84%	81.56%	91.72%	91.37%	91.47%	93.74%	92.96%	93.29%	0,02830
DRCW k=1	85.23%	72.60%	76.33%	86.03%	78.68%	80.91%	92.74%	92.00%	92.32%	91.78%	91.42%	91.51%	4,02127
DRCW k=2	85.99%	72.60%	76.51%	85.94%	78.19%	80.47%	92.36%	91.66%	91.94%	91.88%	91.48%	91.56%	4,02127
DRCW k=3	85.68%	72.47%	76.36%	86.45%	78.80%	81.08%	92.24%	91.48%	91.79%	92.38%	91.81%	91.99%	4,02127
DRCW k=4	85.62%	72.54%	76.40%	86.13%	78.76%	80.97%	92.09%	91.33%	91.65%	92.83%	91.93%	92.26%	4,02127

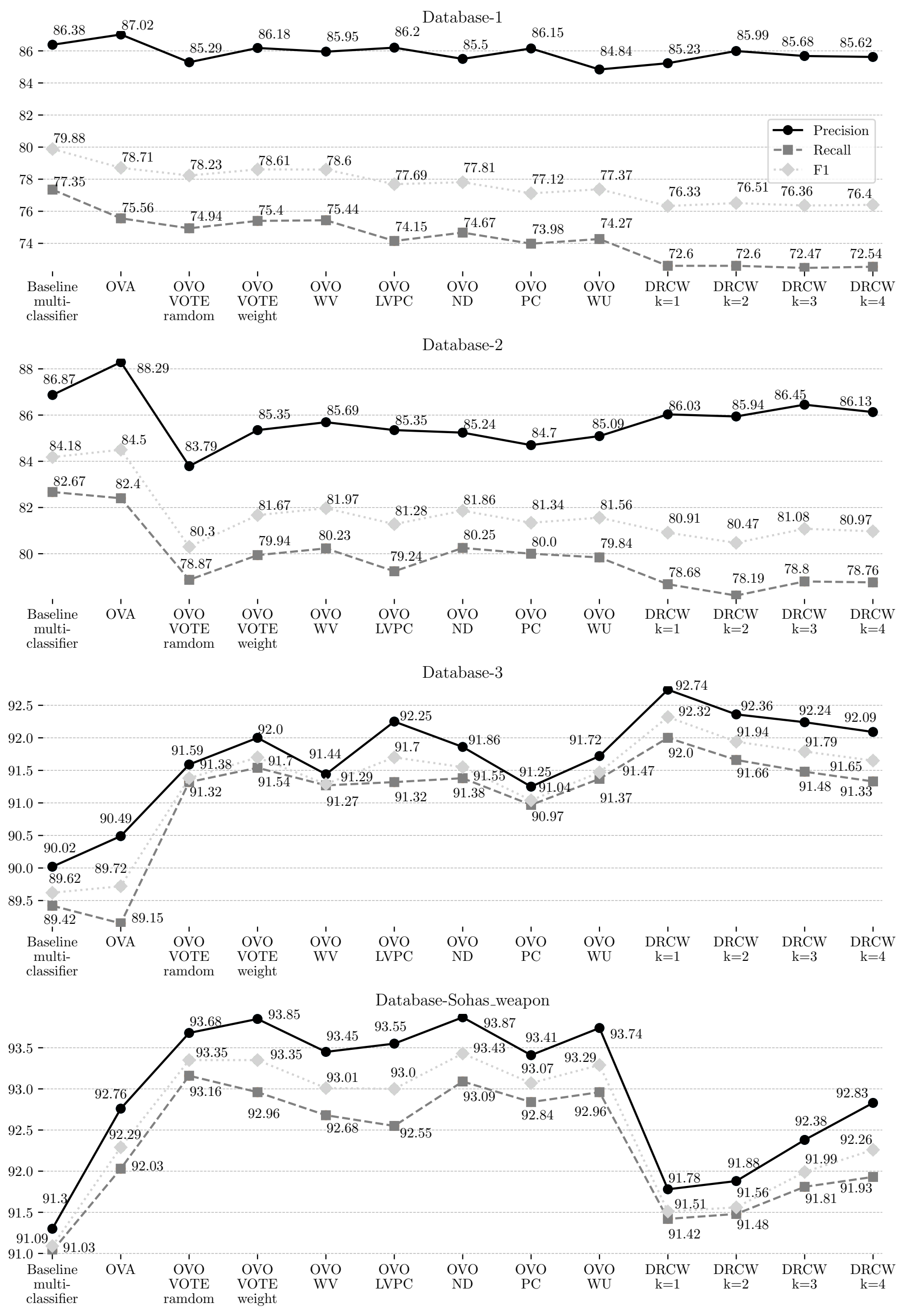

Figure 5: Results of each classification approach trained on Database-1, 2, 3 and Sohas_weapon and tested on Database-Sohas_weapon-Test.

As it can be observed from Figure 5, in general, the performance of all the approaches increases from Database-1 to Database-Sohas_weapon. In particular, when trained on Database-1, OVA and OVO provide similar performance as the baseline multi-classifier. On Database-2, OVA obtains the best performance over all the methods. On Database-3 and Database-Sohas_weapon, all the OVO aggregation methods provide better performance than the baseline multi-classifier.

DRCW-OVO with k=1 gets the best results on Database-3. On Database-Sohas_weapon, OVO ND provide the best results with a precision of 93,87%, recall of 93,09% and F1 of 93,43%. The improvement with respect to the baseline multi-classifier on Database-Sohas_weapon are 2,57% in precision, 2,06% in recall and 2,34% in F1.

However, in terms of execution time, DRCW-OVO takes 4,04 seconds per frame as it calculates the distance between all the images in the database. This makes DRCW-OVO inappropriate for real time processing. Therefore, for evaluating our proposal, we selected only the models that provide a good accuracy/execution time trade-off, OVO with different aggregation rules, VOTE random, VOTE by weight, WV, LVPC, ND, PC and PE.

As conclusion of this evaluation, the use of binarization techniques produces better results than the baseline multi-classifier. Besides, this kind of techniques could be used in real time.

4.2. Similar objects: Analysis

The purpose of this subsection is to study the behaviour of the binarization techniques in harder problems when the similarity between all the objects of the database is higher such as in the case of smartphone, bill, purse and card. To this end, we eliminated the pistol and knife class from Database-Sohas_weapon and Database-Sohas_weapon-Test obtaining Database-Sohas_weapon-Without_Pistol&Knife and Database-Sohas_weapon-Test_Without_Pistol&Knife.

The performance of all the analysed approaches, baseline multi-classifier, OVA and OVO with diverse aggregation rules, VOTE random, VOTE by weight, WV, LVPC, ND, PC, PE, DRCW with k=1, 2, 3 and 4, when trained on Database-Sohas_weapon-Without_Pistol&Knife and tested on Database-Sohas_weapon-Test_Without_Pistol&Knife is provided in Table 4.

Table 4: Result of all the classification approaches trained on Database-Sohas_weapon-Without_Pistol&Knife and tested on Database-Sohas_weapon-Test_Without_Pistol&Knife.

Database-Sohas_weapon-Without_Pistol&Knife			
	Precision	**Recall**	**F1**
Baseline multi-classifier	91,27%	90,46%	90,63%
OVA	91,70%	91,28%	91,32%
OVO VOTE random	92,63%	92,69%	92,62%
OVO VOTE weight	93,51%	93,41%	93,39%
OVO WV	93,29%	93,20%	93,18%
OVO LVPC	93,07%	92,81%	92,87%
OVO ND	93,28%	93,02%	93,08%
OVO PC	93,29%	93,20%	93,18%
OVO PE	93,07%	92,81%	92,87%
DRCW-OVO k=1	**93,76%**	**93,46%**	**93,48%**
DRCW-OVO k=2	93,22%	92,95%	93,03%
DRCW-OVO k=3	93,02%	92,75%	92,82%
DRCW-OVO k=4	93,02%	92,75%	92,82%

As we can observe from Table 4, DRCW-OVO k=1 achieves the best mean values of Precision, Recall and F1, with the respective values of 93,76%, 93,46% and 93,48%. However, these results were similar with respect to the study that includes pistol and knife. For further analysis, Table 5 shows the confusion matrices of the study with and without pistol and knife with its best aggregation method, which obtains the highest performance, OVO-ND.

Table 5: Confusion matrix of best Database-Sohas_weapon and Database-Sohas_weapon-Without_Pistol&Knife, OVO-ND and DRCW-OVO k=1 respectively.

	Database-Sohas_weapon								
	Bill	**Knife**	**Purse**	**Pistol**	**Smartphone**	**Card**	**Precision**	**Recall**	**F1**
Bill	118	1	0	0	0	2	97,52%	95,93%	96,72%
Knife	0	457	2	2	4	0	98,28%	97,23%	97,75%
Purse	3	0	94	3	10	0	85,45%	90,38%	87,85%
Pistol	0	10	4	289	1	3	94,14%	98,30%	96,17%
Smartphone	0	1	4	0	99	1	94,29%	86,09%	90,00%
Card	2	1	0	0	1	58	93,55%	90,63%	92,06%
							93,87%	93,09%	93,43%

	Database-Sohas_weapon-Without_Pistol&Knife								
	Bill	**Knife**	**Purse**	**Pistol**	**Smartphone**	**Card**	**Precision**	**Recall**	**F1**
Bill	120	-	1	-	0	2	97,56%	97,56%	97,56%
Purse	2	-	101	-	13	0	87,07%	97,12%	91,82%
Smartphone	0	-	2	-	100	3	95,24%	86,96%	90,91%
Card	1	-	0	-	2	59	95,16%	92,19%	93,65%
							93,76%	93,46%	93,48%

As it can be observed from Table 5, the mean precision, recall and F1 have a similar value with respect to the case with pistol and knife due to the fact that the pistol and knife class achieves the highest performance over the rest of objects. This can be explained by the unbalance of the database and also by the high quality and quantity of the pistol and knife images.

DRCW-OVO with k=1 trained on Database-Sohas_weapon-Without_Pistol&Knife commits less errors on most similar objects, bill, purse, smartphone and card. As summary, the binarization approach in general and DRCW-OVO with k=1 in particular helps differentiating correctly between similar objects.

This study of similar objects shows how the binarization techniques increase the performance in difficult situations where OVO with multiple aggregation methods obtain the highest performance.

4.3. Evaluation of ODeBiC methodology on surveillance videos

In this section we analyse the methodology ODeBiC using four surveillance videos described on Table 2.

For training the detection models, we used Database-Sohas_weapon-Detection whose characteristics are summarised in Table 1. We consider in this analysis only the classification models that are appropriate for real time execution, OVO with different aggregation rules, VOTE random, VOTE by weight, WV, LVPC, ND, PC and PE.

For the first step of ODeBiC methodology, we used Faster-RCNN based on ResNet-101 trained on Database-Sohas_weapon-Detection. At this stage the detection model analyses the videos frame by frame and outputs the boxes with a detection confidence higher than a minimum threshold. These boxes will be analysed by a binarization method at the second stage by OVO techniques. For comparison purposes we used Faster-RCNN based on ResNet-101 as baseline detector.

Table 6 shows the performance of ODeBiC methodology when using different thresholds, 10%, 50%, 70% and 90%, in the first level of ODeBiC methodology. The threshold refers to the confidence of the model in detecting the considered objects. For the second level, we considered OVO binarization technique with different aggregation methods, vote random, vote weight, WV, LVPC, ND, PC and PE. The actual number of pistols, knives and similar objects in each video is indicated as number of GT (Ground Truth).

Table 6: Results of ODeBiC methodology on four surveillance videos.

		Threshold 10%			Threshold 50%			Threshold 70%			Threshold 90%		
		TP	**FP**	**Precision**	**TP**	**FP**	**Precision**	**TP**	**FP**	**Precision**	**TP**	**FP**	**Precision**
Video 1 1776 GT	Baseline	1189	630	65,37%	994	346	74,18%	926	272	77,30%	843	177	82,65%
	OVO VOTE Random	1540	279	84,66%	1156	184	86,27%	1037	161	86,56%	900	120	88,24%
	OVO VOTE Weight	1535	284	84,39%	1150	190	85,82%	1035	163	86,39%	898	122	88,04%
	OVO WV	1545	274	**84,94%**	1158	182	**86,42%**	1043	155	**87,06%**	903	117	**88,53%**
	OVO LVPC	1529	290	84,06%	1148	192	85,67%	1037	161	86,56%	900	120	88,24%
	OVO ND	1535	284	84,39%	1150	190	85,82%	1036	162	86,48%	898	122	88,04%
	OVO PC	1525	294	83,84%	1137	203	84,85%	1022	176	85,31%	882	138	86,47%
	OVO PE	1533	286	84,28%	1155	185	86,19%	1037	161	86,56%	899	121	88,14%
Video 2 1995 GT	Baseline	1248	617	66,92%	1064	332	76,22%	992	235	80,85%	870	133	**86,74%**
	OVO VOTE Random	1368	497	73,35%	1084	312	77,65%	972	255	79,22%	821	182	81,85%
	OVO VOTE Weight	1385	480	74,26%	1094	302	78,37%	981	246	79,95%	826	177	82,35%
	OVO WV	1402	463	75,17%	1101	295	78,87%	985	242	80,28%	828	175	82,55%
	OVO LVPC	1367	498	73,30%	1078	318	77,22%	970	257	79,05%	818	185	81,56%
	OVO ND	1380	485	73,99%	1091	305	78,15%	978	249	79,71%	823	180	82,05%
	OVO PC	1480	385	**79,36%**	1148	248	**82,23%**	1022	205	**83,29%**	848	155	84,55%
	OVO PE	1378	487	73,89%	1092	304	78,22%	977	250	79,63%	825	178	82,25%
Video 3 1867 GT	Baseline	1250	557	69,18%	1073	298	78,26%	1014	241	80,80%	901	158	85,08%
	OVO VOTE Random	1403	404	77,64%	1116	255	81,40%	1041	214	82,95%	911	148	86,02%
	OVO VOTE Weight	1417	390	78,42%	1127	244	82,20%	1051	204	83,75%	917	142	**86,59%**
	OVO WV	1421	386	78,64%	1126	245	82,13%	1049	206	83,59%	914	145	86,31%
	OVO LVPC	1406	401	77,81%	1118	253	81,55%	1043	212	83,11%	910	149	85,93%
	OVO ND	1409	398	77,97%	1120	251	81,69%	1045	210	83,27%	913	146	86,21%
	OVO PC	1443	364	**79,86%**	1139	232	**83,08%**	1056	199	**84,14%**	912	147	86,12%
	OVO PE	1407	400	77,86%	1118	253	81,55%	1044	211	83,19%	914	145	86,31%
Video 4 2177 GT	Baseline	1502	742	66,93%	1301	381	77,35%	1211	292	80,57%	1063	208	83,63%
	OVO VOTE Random	1816	428	80,93%	1404	278	83,47%	1266	237	84,23%	1093	178	86,00%
	OVO VOTE Weight	1819	425	81,06%	1409	273	83,77%	1270	233	84,50%	1096	175	86,23%
	OVO WV	1821	423	81,15%	1409	273	83,77%	1269	234	84,43%	1094	177	86,07%
	OVO LVPC	1794	450	79,95%	1392	290	82,76%	1253	250	83,37%	1083	188	85,21%
	OVO ND	1819	425	81,06%	1408	274	83,71%	1269	234	84,43%	1095	176	86,15%
	OVO PC	1874	370	**83,51%**	1440	242	**85,61%**	1296	207	**86,23%**	1113	158	**87,57%**
	OVO PE	1807	437	80,53%	1397	285	83,06%	1260	243	83,83%	1087	184	85,52%

In general, as it can be observed from Table 6, the proposed methodology based on OVO technique with all of the aggregation methods overcomes the baseline detection model in the analysed videos. The best results were achieved by OVO with PC or WV aggregation method in the videos, and with all threshold values.

The results can be summarised as follows:

- ODeBiC methodology based on OVO aggregation method overcomes the baseline model in precision between 10,68% and 19,57% for threshold of 10%, between 4,81% and 12,24% for threshold of 50%, between 2,44% and 9,77% for threshold of 70% and between -2,19% and 5,88% for a threshold of 90%.

- In terms of false positives, it reduces the number of false positives between 34,64% and 56,50% for threshold of 10%, between 22,14% and 47,39% for a threshold of 50%, between 12,76% and 43,01% for a threshold of 70% and between -16,54% and 33,89% for a threshold of 90%.
- In terms of execution time, the baseline detection model takes 0,12341 seconds (equivalent to 8 fps) and ODeBiC methodology with OVO PC takes around 0,16834 (equivalent to 6 fps), which is appropriate for near real time system.

In summary, ODeBiC methodology runs in near real time and achieves an improvement of up to 56,50% using an aggregation method of OVO.

5. Conclusions and future work

This work presents the two level methodology ODeBiC based on deep learning for the detection of small objects that can be handled similarly. We considered as case study the detection of small objects that can be confused with a handgun or a knife in surveillance videos. We built a training database, called Sohas_weapon, which includes six objects that can be confused with a weapon as they are commonly handled in a similar way: pistol, knife, smartphone, bill, purse or card.

Our experiments showed that ODeBiC methodology based on an aggregation method of OVO reduced the number of false positives by up to 56,50% and between a -2,19% and 19,57% in precision, depending on the threshold, with respect to the baseline detection model.

ODeBiC methodology can be used as a detection model in surveillance videos as it produces robust output, considerably reduces the number of false positives and obtains better precision than the baseline detection model.

As future work, we will design a new pre-processing strategy to filter noisy instances that can cause confusion in the CNN model.

Acknowledgements

This research work is partially supported by the Spanish Ministry of Science and Technology under the project TIN2017-89517-P and the project DeepSCOP-Ayudas Fundación BBVA a Equipos de Investigación Científica en Big Data 2018. Siham Tabik was supported by the Ramon y Cajal Programme (RYC-2015-18136). The Titan X Pascal used for this research was donated by the NVIDIA Corporation.

References

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: a system for large-scale machine learning., in: OSDI, pp. 265–283.

[2] Abe, S., 2003. Analysis of multiclass support vector machines. Thyroid 21, 3772.

[3] Anand, R., Mehrotra, K., Mohan, C.K., Ranka, S., 1995. Efficient classification for multiclass problems using modular neural networks. IEEE Transactions on Neural Networks 6, 117–124.

[4] Castillo, A., Tabik, S., Pérez, F., Olmos, R., Herrera, F., 2019. Brightness guided preprocessing for automatic cold steel weapon detection in surveillance videos with deep learning. Neurocomputing 330, 151–161.

[5] Chen, X., Fang, T., Huo, H., Li, D., 2015. Measuring the effectiveness of various features for thematic information extraction from very high resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing 53, 4837–4851.

[6] Clark, P., Boswell, R., 1991. Rule induction with cn2: Some recent improvements, in: European Working Session on Learning, Springer. pp. 151–163.

[7] Cruz, R.M., Sabourin, R., Cavalcanti, G.D., 2018. Dynamic classifier selection: Recent advances and perspectives. Information Fusion 41, 195–216.

[8] Dai, K.J., R-FCN, Y.L., 2016. Object detection via region-based fully convolutional networks. arxiv preprint. arXiv preprint arXiv:1605.06409 .

[9] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, Ieee. pp. 248–255.

[10] Fernández, A., Calderón, M., Barrenechea, E., Bustince, H., Herrera, F., 2009. Enhancing fuzzy rule based systems in multi-classification using pairwise coupling with preference relations. EUROFUSE 9, 39–46.

[11] Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F., 2018. Learning from Imbalanced Data Sets. Springer.

[12] Friedman, J.H., 1996. Another approach to polychotomous classification. Technical Report, Statistics Department, Stanford University .

[13] Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F., 2011. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recognition 44, 1761–1776.

[14] Galar, M., Fernández, A., Barrenechea, E., Herrera, F., 2015. Drcw-ovo: distance-based relative competence weighting combination for one-vs-one strategy in multi-class problems. Pattern recognition 48, 28–42.

[15] Hastie, T., Tibshirani, R., 1998. Classification by pairwise coupling, in: Advances in neural information processing systems, pp. 507–513.

[16] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

[17] Hu, J., Shen, L., Sun, G., 2017. Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 7.

[18] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., 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.

[19] Huhn, J.C., Hullermeier, E., 2009. Fr3: A fuzzy rule learner for inducing reliable classifiers. IEEE Transactions on Fuzzy Systems 17, 138.

[20] Hüllermeier, E., Brinker, K., 2008. Learning valued preference structures for solving classification problems. Fuzzy Sets and Systems 159, 2337–2352.

[21] Knerr, S., Personnaz, L., Dreyfus, G., 1990. Single-layer learning revisited: a stepwise procedure for building and training a neural network, in: Neurocomputing. Springer, pp. 41–50.

[22] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

[23] Lei, H., Govindaraju, V., 2005. Half-against-half multi-class support vector machines, in: International Workshop on Multiple Classifier Systems, Springer. pp. 156–164.

[24] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755.

[25] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector, in: European conference on computer vision, Springer. pp. 21–37.

[26] Olmos, R., Tabik, S., Castillo, A., Pérez, F., Herrera, F., 2019. A binocular image fusion approach for minimizing false positives in handgun detection with deep learning. Information Fusion 49, 271 – 280.

[27] Olmos, R., Tabik, S., Herrera, F., 2018. Automatic handgun detection alarm in videos using deep learning. Neurocomputing 275, 66–72.

[28] Orlovsky, S., 1978. Decision-making with a fuzzy preference relation. Fuzzy sets and systems 1, 155–167.

[29] Öztürk, K., Yilmaz, M.B., 2017. A comparison of classification approaches for deep face recognition, in: Computer Science and Engineering (UBMK), 2017 International Conference on, IEEE. pp. 227–232.

[30] Pathak, A.R., Pandey, M., Rautaray, S., 2018. Application of deep learning for object detection. Procedia computer science 132, 1706–1717.

[31] Pele, O., Werman, M., 2010. The quadratic-chi histogram distance family, in: European conference on computer vision, Springer. pp. 749–762.

[32] Platt, J.C., Cristianini, N., Shawe-Taylor, J., 2000. Large margin dags for multiclass classification, in: Advances in neural information processing systems, pp. 547–553.

[33] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.

[34] Rocha, A., Goldenstein, S.K., 2014. Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches. IEEE Transactions on Neural Networks and Learning Systems 25, 289–302.

[35] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115, 211–252.

[36] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

[37] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.

[38] Wu, T.F., Lin, C.J., Weng, R.C., 2004. Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005.

[39] Yu, M., Gong, L., Kollias, S., 2017. Computer vision based fall detection by a convolutional neural network, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ACM. pp. 416–420.

[40] Yuan, J., Hou, X., Xiao, Y., Cao, D., Guan, W., Nie, L., 2019. Multi-criteria active deep learning for image classification. Knowledge-Based Systems 172, 86–94.

[41] Zhang, C., Bi, J., Xu, S., Ramentol, E., Fan, G., Qiao, B., Fujita, H., 2019. Multi-imbalance: An open-source software for multi-class imbalance learning. Knowledge-Based Systems .

[42] Zhang, Q., Yang, L.T., Chen, Z., Li, P., 2018. A survey on deep learning for big data. Information Fusion 42, 146–157.