Information Fusion 49 (2019) 271–280

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full Length Article

A binocular image fusion approach for minimizing false positives in handgun detection with deep learning

Roberto Olmos, Siham Tabik*, Alberto Lamas, Francisco Pérez-Hernández, Francisco Herrera

Andalusian Research Institute in Data Science and Computational Intelligence, University of Granada, Granada, 18071, Spain

ARTICLE INFO

Keywords:
Classification
Detection
Deep learning
Convolutional neural networks (CNNs)
Faster R-CNN
VGG-16
Selective research
Region proposals

ABSTRACT

Object detection models have known important improvements in the recent years. The state-of-the art detectors are end-to-end Convolutional Neural Network based models that reach good mean average precisions, around 73%, on benchmarks of high quality images. However, these models still produce a large number of false positives in low quality videos such as, surveillance videos. This paper proposes a novel image fusion approach to make the detection model focus on the area of interest where the action is more likely to happen in the scene. We propose building a low cost symmetric dual camera system to compute the disparity map and exploit this information to improve the selection of candidate regions from the input frames. From our results, the proposed approach not only reduces the number of false positives but also improves the overall performance of the detection model which make it appropriate for object detection in surveillance videos.

1. Introduction

The early detection of potentially violent situations is of paramount importance for citizens security. One way to predict these situations is by detecting the presence of dangerous objects such as handguns in surveillance videos. Current surveillance and control systems still require the supervision and intervention of humans. One innovative solution to this problem is to equip surveillance or control cameras with an accurate automatic handgun detection alert system.

Detecting handguns in monitored indoor areas such as, jewelries and banks, is a complex task because surveillance videos often have low quality, some areas of the frames maybe distorted and blurry, which makes their identification difficult even for the human eye. On the other hand, a smart surveillance system designed for violence prevention must produce an alarm in near-real time and only when it is completely confident about the presence of a handgun in the scene. In real world surveillance scenarios, the alarms invoked by false positives must be reduced to the minimum.

The state-of-the art object detection models, based on Convolutional Neural Networks (CNNs), are showing promising results on the 200 objects benchmark in the ILSVRC (Large Scale Visual Recognition Challenge) competitions [17]. For instance, the most accurate detection model in ILSVRC-2017 reached a mean accuracy of around 73%, which is impressive even for a high quality images benchmark.

Achieving good accuracies with the state-of-the art detection models on surveillance videos is still an open issue. As far as we know, the first and unique related work in this context, presented in [14], showed good accuracies on monocular Youtube videos of movies from the nineties but produced an important number of false positives. Most false positives are objects from the background produced by the fact that the selection search methods used in the state-of-the art detection models assume that all the areas of the input image are candidate regions [6,22].

This work proposes a novel binocular image fusion approach for reducing the number of false positives in the detection of handguns in surveillance videos. In particular, we calculate the disparity map based on a symmetric binocular vision approach and use this information to preselect the areas of interest where the action is more likely to happen in the scenario. As far as we know, this work is the first in employing symmetric binocular vision and disparity maps to reduce the number of false positives in object detection in videos. We focus on RGB surveillance videos recorded in indoor public places such as, jewelries or banks, where commonly actions such as robberies may occur under artificial light.

The main contributions of this work are:

- Build a simple binocular system based on two symmetric IP surveillance cameras to compute the depth of pixels in the scene.
- Propose a novel binocular image fusion approach for reducing the number of false positives in the detection of handguns with deep learning models.
- Present the first work in addressing the detection of handguns in the field of video surveillance.

* Corresponding author.
E-mail address: siham@ugr.es (S. Tabik).

https://doi.org/10.1016/j.inffus.2018.11.015
Received 15 January 2018; Received in revised form 15 November 2018; Accepted 28 November 2018
Available online 29 November 2018
1566-2535/© 2018 Elsevier B.V. All rights reserved.

This paper is organized as follows. Section 2 gives a brief analysis of the most related works to our work. Section 3 reviews the stat-of-the art CNNs and transfer-learning. Section 4 described the image fusion approach proposed in this work. Section 5 provides and analyzes the obtained results and finally Section 6 conclusions.

2. Related works

Using the disparity information for improving handguns detection in surveillance videos is related in part to two research areas. The first area uses symmetric or asymmetric dual cameras to improve the quality of the image or perform a three-dimensional reconstruction of the scene. The second area focuses on improving the detection of a large number of common objects in images using CNNs based models.

2.1. Binocular vision based on asymmetric or symmetric dual cameras

Binocular vision based on asymmetric or symmetric dual cameras provides important information than cannot be obtained by monocular cameras. Asymmetric dual camera systems, where each camera has a different field of view and resolution, are often used to enhance the resolution [9], digital zooming [13,23] or overall quality of the image [7]. While, symmetric dual camera systems allow calculating the disparity and depth information to improve the recognition or localization of objects. For example, the authors in [24] proposed an ensemble of pedestrian detection model as follows. They first run their detector on the left and right images individually, join the obtained bounding boxes in one single image and re-run the detector on the third joined image. The result of the detection is calculated as the score fusion of the three detections. The authors in [8] proposed generating a 3D model to calculate the exact 3D position of the lesion in image-guided operations, in neurosurgery.

Our work is different from these works in that we propose using the disparity information obtained from a symmetric dual camera to improve the set of region proposals with the objective of reducing the number of false positives in hand-gun detection in surveillance videos. The proposed symmetric binocular vision approach is similar to the human binocular vision system, it has the capacity of detecting the background based on the perceived dimensional data, which make it much more appropriate than the monocular pixel intensity methods.

2.2. Detection models using CNNs

The most accurate detection models reformulate the problem of object detection into a combination of a search selection technique with a CNN classification model. The selection method generates candidate-regions from the input image, commonly called region proposals, then a CNN-based classifier is run on each one of these regions. Currently, the most influential detection model is Faster R-CNN which is an evolution of its predecessors R-CNN and Fast R-CNN as follows:

- R-CNN [3] was the first detection models that included a CNN-based classifier. It uses an external box generator to produce several crops from the image, on the order of 2000 boxes called region proposals. It then runs VGG-based classifier on each one of these region proposals. Afterwards, the output of the CNN is feeded to two predictors: i) a SVM classifier that predicts the class of each region and ii) a linear regressor that predicts the bounding box of the detected class. R-CNN provides good performance on the well know PASCAL-VOC however it is too slow due to the high redundant computation since the CNN is applied on 2000 crops.
- Fast R-CNN [2] improves the speed of R-CNN by reorganizing the calculation as follows. It first extracts the features of the entire input image before generating the region proposals. It replaces the SVM classifier with a softmax layer so that the CNN is extended to directly calculate the prediction of the object-class. The remaining bottleneck in Fast R-CNN was the selective search technique for generating the region proposals.
- Faster R-CNN [16], its main difference with respect to Fast R-CNN was converting the selective search algorithm into a faster network called Region Proposal Network (RPN). This change improved both, accuracy and speed. Faster R-CNN can be seen as a meta-architecture that combines a RPN with a feature extractor, e.g., VGG-16. Although recent meta-architectures such as R-FCN [1] and SSD [12] achieve better speeds, Faster R-CNN still provide the best accuracy [6].

The first work that proposed detecting handgun detection in videos using CNN-based models was [14]. The model reached good accuracies but only on monocular Youtube videos using Faster R-CNN. The evaluation of the detection results provided in [14] did not consider real world surveillance videos. The present work proposes improving this model for real surveillance videos.

3. CNN-based detection models

The state-of-the art models address the detection problem by reformulating it into a classification problem, where the CNN-classifier is evaluated over a set of candidate regions extracted from the input image. In this work, we considered the most accurate detection meta-architecture, Faster R-CNN based on four feature extractors, VGG-16 [19], ResNet [4], Inception-ResNet-v2 [20] and NAS [25]. We also consider transfer-learning to improve the learning of the evaluated CNNs models.

3.1. CNNs classification models

VGGNet was the first runner-up in ILSVRC 2014 [19]. It was used to show that the depth of the network is critical to the performance. The largest VGGNet architecture, VGG-16, involves 144 million parameters from 16 convolutional layers with very small receptive fields 3×3, five max-pooling layers of size 2×2, three fully-connected layers, and a linear layer with Softmax activation in the output. This model also uses dropout regularization in the fully-connected layer and applies ReLU activation to all the convolutional layers.

ResNet won the first place on the ILSVRC 2015 and is currently the most accurate and deepest CNN architecture [4]. It has 152 layers and 25.5 million parameters. Its main characteristic with respect to the previous deep CNNs, e.g., GoogLeNet [21], is that ResNet creates multiple paths through the network within each residual module.

Deep CNNs, such as VGG and ResNet, are generally trained based on the prediction loss minimization. Let x and y be the input images and corresponding output class labels, the objective of the training is to iteratively minimize the average loss defined as

$$J(w) = \frac{1}{N}\sum_{i=1}^{N} L(f(w; x_i), y_i) + \lambda R(w) \quad (1)$$

This loss function measures how different is the output of the final layer from the ground truth. N is the number of data instances (mini-batch) in every iteration, L is the loss function, f is the predicted output of the network depending on the current weights w, and R is the weight decay with the Lagrange multiplier λ. It is worth mentioning that in the case of GoogLeNet, the losses of the two auxiliary classifiers are weighted by 0.3 and added to the total loss of each training iteration. The Stochastic Gradient Descent (SGD) is commonly used to update the weights.

$$w_{t+1} = \mu w_t - \alpha \Delta J(w_t) \quad (2)$$

where μ is the momentum weight for the current weights w_t and α is the learning rate.

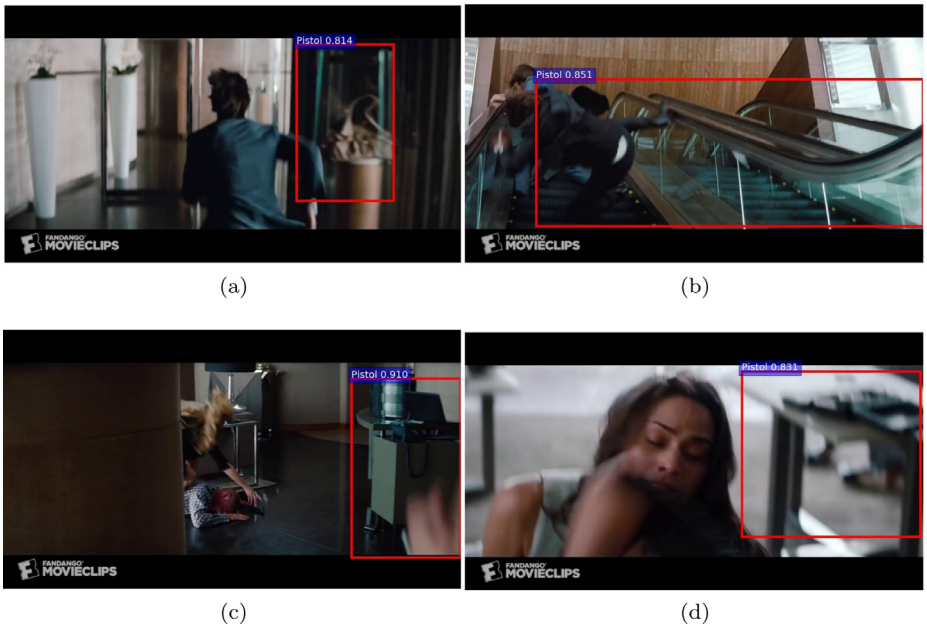


Fig. 1. Examples of false positives detected in the background, (a) a handbag in a showcase, (b) the mechanical stairs, (c) a screen and a laptop under a table and (d) an open case under a table.

The network weights, w_t, can be randomly initialized if the network is trained from scratch. However, this is suitable only when a large labeled training-set is available, which is expensive in practice. Several previous studies have shown that transfer-learning [18] can help overcoming this limitation.

3.2. Transfer learning and fine-tuning

Transfer learning consists of re-utilizing the knowledge learnt from one problem to another related one [15]. In practice, it is applied by initializing the weights of the network, w_t in Eq. (2), with the pre-trained weights.

Re-training or fine-tuning the entire network (i.e., updating all the weights) is only used when the new dataset is large enough, otherwise, the model could suffer overfitting especially among the first layers of the network. Since these layers extract low-level features, e.g., edges and color, they do not change significantly and can be utilized for several visual recognition tasks. The last learnable layers of the CNN are gradually adjusted to the particularities of the problem to extract high level features of the new database.

We analyzed four networks, VGG-16, Inception-Resnet v2, NAS and ResNet. We initialized the weights of VGG-16 with the pre-trained weights of the same architectures on ImageNet dataset (around 1.28 million images over 1000 generic object classes) [10] and initialized Inception-Resnet v2, NAS and ResNet using the pre-trained weights on COCO dataset [11] (around 328k images over 91 objects types).

The fine-tuning process applied in this work is as follows:

1. We remove the last pooling layer and the fully connected layers from the CNN classifier.
2. We add the RoI pooling layer, the Region Proposal Network and the fully connected layers that will calculate the final prediction.
3. Afterwards, we initialize the CNN using the pre-trained weights and the rest of the weights with random numbers.
4. Finally, we fine-tune the weights of all the network during 50,000 epochs and update the learning rate progressively. This process allows the network to tune the predictions of the classification and the detection of the boxes to the particularities of the handgun detection problem.

4. The proposed image fusion approach

We reformulate the problem of handgun detection into a two class detection problem, where the positive class is pistol. In the previous study [14], the authors addressed the problem of handgun detection by applying the detection to a stream of images obtained directly from the video. While this method proved to be efficient for real time detection, the results still show some issues in specific situations.

The most frequent false positives were located in the background as illustrated by the examples provided in Fig. 1. As we can observe from these images the detector considers as pistol, with a high probability, objects in the background that contain common features with the pistol. For example, the objects of dark metallic colors with a silhouette of a pistol shown in Fig. 1(b) and (d) are identified as pistol by the detector. This is due to the fact that the region proposals generators used in the state-of-the art detection model assume that all the regions of the input image are potential candidates of pistol. However, in real-world scenarios, the pistol can be located only in a limited area of the frame. The background of the scene could contain several objects that could produce a high number of false positives and make the search algorithm not focus on the area of interest.

We propose reducing the number of false positives by using the disparity information in six steps as depicted in Fig. 2:

1. The frames are obtained from two symmetric commercial surveillance IP cameras.
2. The disparity map is calculated based on the binocular frames obtained from the cameras.
3. The background objects are eliminated.
4. A pre-selection of the areas of interest is performed.
5. The obtained mask is applied to one of the original frames.
6. Finally, the detection process is applied.

In the following sections we analyze each of these steps.

4.1. Frames acquisition from the symmetric dual camera: system setup

In general, dual camera systems need an external switch to synchronize the time at which the frames are captured. However, our purpose in this work is to build a system based only on consumer cameras without

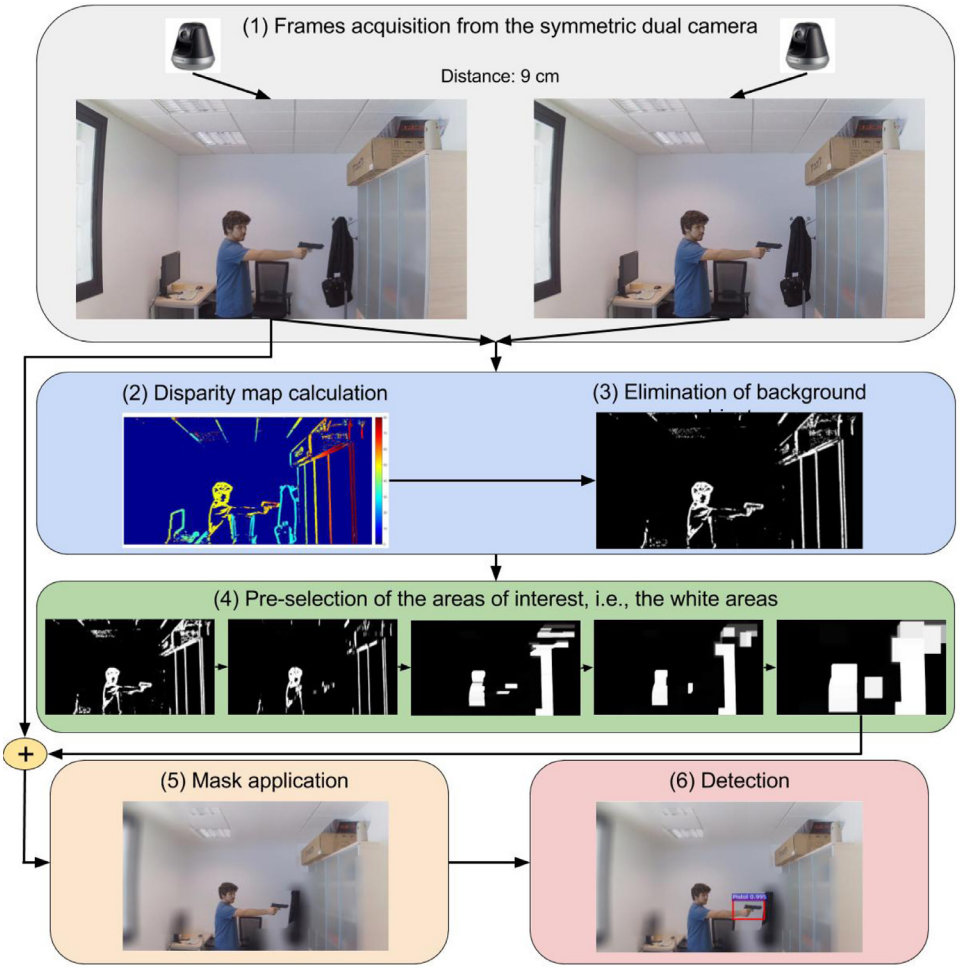


Fig. 2. Flowchart of our approach. During the test/evaluation stage, first, (1) the frames are obtained from a symmetric dual camera system, (2) the disparity map is calculated, (3) the background objects are eliminated, (4) the pre-selection of the areas of interest is performed, (5) the obtained mask is applied to the frame and finally, (6) the detection process is performed.

switch. We used the next setup, the central axes of the field of vision of the cameras are set in parallel and the distance between the centers of the two camera lens is set to 9 cm. The dual camera system was calibrated using the image of a chessboard of 2.4 cm × 2.4 cm squares.

4.2. Disparity map calculation

To calculate the disparity map, we evaluated two algorithms, the Block Matching (BM) Algorithm and its variation, the Semi-Global Block Matching (SGBM) Algorithm [5].

In general, these algorithms calculate the distance between an area of pixels in the image taken by the left camera and its corresponding area of pixels in the image taken by the right camera. The closer an object is from the camera, the larger the distance between the object and its protection. If the object is farther from the cameras, the distance between the object and its projection will be smaller or null for very far objects.

The BM and SGBM algorithms use this information to estimate the distance between the camera and the objects in the scene. The result of this estimation is represented in the form of a disparity map as illustrated in Fig. 3(c) and (d).

By analyzing both algorithms on our surveillance videos, we found that the BM algorithm is too sensitive to changes in the environment (e.g., light, camera angle) and scenario conditions (e.g., distance between the objects in the scene). Reducing the size of the used blocks to improve the stability of BM produces discontinuous and imprecise disparity maps as shown in Fig. 3(c) and (d). Whereas, the SGBM algorithm shows a higher stability, more continuous and precise disparity detection than BM, and provides the necessary dimensional information to differentiate the objects in the background, as it can be seen in Fig. 3(d).

4.3. Elimination of background objects

After calculating the disparity map, we chose a limit distance to eliminate the objects situated behind that distance. The selection of the limit distance in the scene depends on the dimensions of the considered scenario, e.g., a room, and on the area where the action occurs. The distance is calculated from the dual camera system.

4.4. Pre-selection of the areas of interest

The binocular disparity map allows distinguishing farther and closer objects from the camera and consequently the objects can be eliminated based on their distance. However, in practice, the produced disparity map has imperfections produced by the effects of the lights in the scene and by the low quality of the original images, which makes the elimination process difficult.

The resulting disparity map contains a certain level of noise and shows discontinuities in several borders, as it can be seen in Fig. 2(4). To clean and improve the segments, we apply a series of morphological binary operations as follows:

- Step1: To preserve the small details in the disparity map, we first apply a uniform dilatation, in both x and y axes, to the white areas in the image.
- Step 2: Iteratively, we apply an erosion process followed by a dilatation process to the obtained image. The dilatation operation

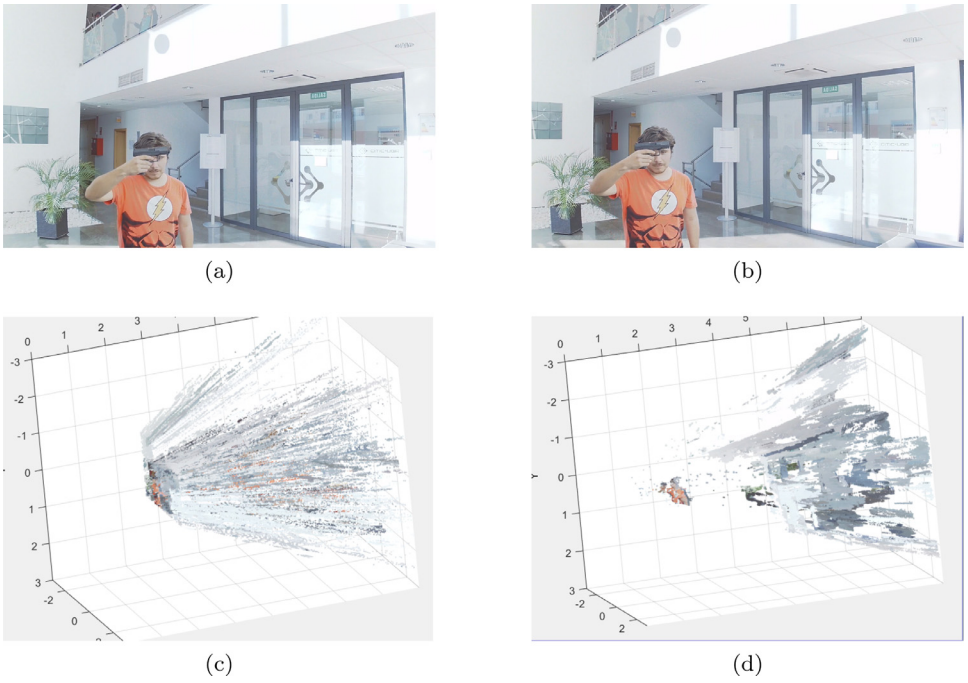


Fig. 3. The disparity map calculated by BM algorithm (c) and SGBM algorithm (d) based on the information obtained from the binocular left (a) and right (b) image.

transforms the lines that are part of a same object into a larger white area. While, the erosion process erases the areas of the image that did not formed a large white region.
This step is applied to different shapes to ensure the elimination of small objects, different kinds of noise and imperfections. At the end of this process, we obtain a mask that indicates the areas of interest.

4.5. Mask application and detection process

The obtained mask from the previous step is applied to the original image as follows. The parts of the original image that correspond to the white areas of the mask are maintained. While the parts of the original image that correspond to the black areas of the mask are blurred. Afterwards, we apply the detector on the entire image although the detector will focus only on the area of interest.

It is worth noting that we have also analyzed eliminating the background objects completely from the original image. We found that complectly eliminating the background maintains the focus of the detector on the area of interest; however, the high contrast between the erased background and the image led the detector to misinterpret the shown area as a big object. To overcome this issue, we used the blurring method to produce a lower contrast between the area of interest and background and consequently prevent from misinterpretations and preserve the capability of the search algorithm to focus only on the areas of interest.

We evaluated four detection models, Faster R-CNN based on VGG-16 with transfer learning from ImageNet and Faster R-CNN based on ResNet, Inception ResNet V2 and NAS with transfer learning from COCO.

5. Experiments and results

The impact of the proposed image fusion approach on the handgun detection model is analyzed in this section. First, we describe the two symmetric cameras systems used in the analysis. Second, we select the best performing detection model by evaluating Faster R-CNN with four feature extractors, VGG-16 with transfer-learning from ImageNet, Inception ResNet v2, NAS and ResNet with transfer-learning from COCO. Then, we analyze and compare the impact of our image fusion approach on the performance of the selected detection model using nine surveillance videos. Finally, we analyze several false positives that were eliminated by applying the proposed image fusion approach.

5.1. Experimental setup

For training the analyzed detection models, we used a dataset that contains 3000 handgun images. This training dataset was built by the authors of this manuscript and it is available through this link http://sci2s.ugr.es/weapons-detection. For testing the impact of the proposed apporach on the performance of the detection, we built nine diverse and challenging test videos using two different systems.

We built two symmetric camera systems using two different commercial cameras as follows:

- System 1: based on a symmetric dual HD IP cameras setup, LOGITECH c525, of resolution 1280×720, maximum frame-rate 30 fps, angular field of view 69° and video Compression Format, H.264 and Motion JPEG.
- System 2: based on a symmetric dual Full HD IP cameras setup, SAMSUNG SNH-V6410PN, of resolution 1920×1080, maximum frame-rate 30 fps, angular field of view 96.1° and video Compression Format, H.264 and Motion JPEG.

Notice that the commercial camera used in system 2 has better properties, resolution and field of view than the one used in system 1. To synchronize the acquisition of images in each system without switch, we limited the maximum frame rate of both cameras to 1 fps.

We recorded nine surveillance videos of a person handling one or two guns in different environments. The first four videos, vid1, vid2, vid3 and vid4, were recorded using system 1, in a challenging house scenarios with multiple and diverse objects in the background. In particular, vid1 and vid4 include a 2d printed image of a gun to test the capability of the system to distinguish between real and fake 2D representation of a gun. vid4 were recorded in a larger space to study the effect of larger distances between the cameras and the object of interest.

The last five videos, vid5, vid6, vid7, vid8 and vid9, were recorded using system 2 in two different scenarios. vid5, vid6 and vid7 were recorded in an office with simple background and with a low probability for producing a high number of concurrent false positives. While,

Table 1
The performance of the handgun detection model using Faster R-CNN with VGG-16 (with transfer-learning from ImageNet), ResNet, Inception Resnet v2 and NAS (with transfer-learning from COCO) on vid6.

	Fusion	# images	TP	FP	TN	FN	Accuracy	Recall	Precision	F1
VGG-16 +	without	124	76	8	28	19	79.39%	80.00%	90.48%	84.92%
ImageNet	with	124	76	**6**	28	19	**80.62%**	80.00%	**92.68%**	**85.88%**
ResNet +	without	124	74	161	0	21	28.91%	**77.89%**	31.49%	44.85%
COCO	with	124	70	**29**	25	25	**63.76%**	73.68%	**70.71%**	**72.16%**
Inception ResNet +	without	124	78	17	29	17	75.89	82.10%	82.10%	82.10%
COCO	with	124	79	**14**	29	16	**78.26%**	**83.15%**	**84.94%**	**84.04%**
NAS +	without	124	65	34	26	30	58.71%	**68.42%**	65.66%	67.01%
COCO	with	124	63	**26**	25	32	**60.27%**	66.31%	**70.78%**	**68.48%**

vid8 and vid9, were recorded in a more realistic surveillance scenario, the entrance of a building with several persons going in and out. The videos from vid1 to vid7 are used for an individual examination of the proposed approach while vid8 and vid9, are used for a global evaluation of the proposed approach.

It is worth to mention that the high reflectance of natural sun light in the second scenario produces some loss of information in some frames. Which makes the calculation of the disparity map difficult. To experimentally evaluate the effectiveness of the proposed fusion technique, we selected the scenes that contain frames with lower reflectance effects.

- vid1: recorded in a house environment with diverse objects in the background, a mirror and several pieces of furniture, under artificial light. The person moves a gun and compares it with a fake 2D gun representation.
- vid2: recorded in the same house environment as vid1. The person holds two guns, a pistol and a revolver, and moves them from left to right then put the revolver in his pocket.
- vid3: recorded in the same house environment. The person moves the pistol while rotating it and partially occluding different of its parts.
- vid4: recorded in a different house environment. This video was recorded in a larger space to study the effect of larger distances between the cameras and the object of interest. The person holds a handgun and a printed image of a gun in different poses. This video is used to test the ability of the system to detect a subject that starts close from the camera and walks away until he is 10 *meters* from the camera.
- vid5: recorded in an office with simple background. The person is moving from closer to further from the camera and pointing with one pistol to different angles in the room.
- vid6: recorded in an office with simple background. The person is moving from left to right and around pointing with the pistol in different angles in the room.
- vid7: recorded in an office with simple background. The person is moving in all directions, first, pointing to all directions with one pistol using one hand and then pointing with two pistols, one in each hand.
- vid8: recorded in the entrance of a building with more complex background. The person is moving from left to right and around pointing with the pistol to different directions.
- vid9: recorded in the entrance of a building with more complex background. The person is moving from left to right and around pointing with two pistols, one pistol in each hand, to different directions.

For the evaluation and comparisons, we used four metrics, *accuracy, precision* (also called positive predictive value, i.e., how many detected pistols are true), *recall* (also known as sensitivity, i.e., how many actual pistols were detected), and *F1 measure*, which evaluates the balance between *precision* and *recall*. Where

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ images},$$

$$precision = \frac{True\ Positives}{True\ Positives + False\ Positives},$$

$$recall = \frac{True\ Positives}{True\ Positives + False\ Negatives},$$

and

$$F1\ measure = 2 \times \frac{precision \times recall}{precision + recall}$$

Where, True Positives (TP) refers to the number of pistols correctly detected in the frames of the input video. For example, if a given frame has two visible pistols, the detection model must produce two bounding boxes that will be considered as two TP only if each bounding-box has an overlapping with the pistol area larger than 70%. False Positives (FP) refer to the number of bounding boxes produced by the detection model in which there is no pistol or a tiny part of the pistol. If a pistol is detected more than once, the number of redundant bounding boxes will be considered as FP, since we have already applied the Non-maximum suppression method to unify redundant detection of the same object. True Negatives (TN) refers to the total number of frames where there is neither visible pistols nor false positives. False negatives (FN) refers to the number of visible pistols that are not detected by the object detection algorithm.

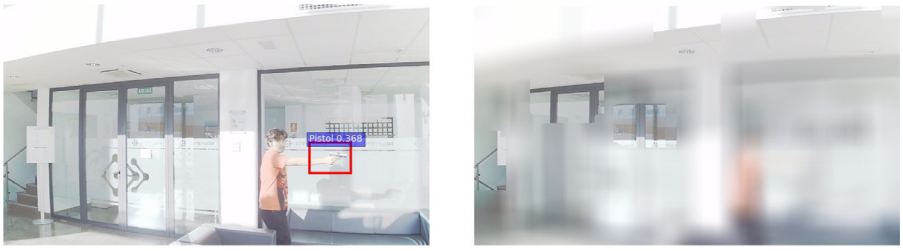
5.2. Transfer learning from imagenet and COCO datasets

Before evaluating the impact of the proposed image fusion approach on the test videos, we first analyzed the performance of Faster R-CNN using four feature extractors, VGG-16 and ResNet, with and without image fusion on vid6. The only pre-trained weights available to us were, the VGG-16 weights pre-trained on ImageNet and Inception Resnet v2, NAS, ResNet weights pre-trained on COCO.

The performance of Faster R-CNN with VGG-16, Inception Resnet v2, NAS, and ResNet with transfer-learning from ImageNet and COCO with and without fusion on vid6 are shown in Table 1. In general, the accuracy, precision, recall and F1_measure of Faster R-CNN(VGG-16 + ImageNet) are much higher than the obtained with Faster R-CNN(ResNet + COCO and NAS + COCO) with and without fusion and higher accuracy, precision and F1 than the obtained with Faster R-CNN(Inception ResNet v2 + COCO). The produced number false positives by Faster R-CNN(VGG-16 + ImageNet) are substantially lower than Faster R-CNN(COCO based models) with and without fusion. This can be explained by the fact that the pre-training acquired from the larger number of classes in ImageNet improves drastically the learning from our handgun dataset available through this link (http://sci2s.ugr.es/weapons-detection). Another factor that could be of importance to comprehend why Faster R-CNN(VGG-16 + ImageNet) outperforms Faster R-CNN(COCO based models) is the presence of the revolver class in ImageNet dataset. This class has a certain level of similarity with our handgun class. COCO does not include any similar object. The low NAS results could be explained by the fact that the training images of our dataset are not all of 1200×1200 pixel or larger as in the original NAS requirement,

Table 2
Performance of the handgun detection model using a monocular video (i.e., obtained from one camera) labeled in the table as *without* and using the proposed binocular fusion technique (i.e., obtained from a symmetric dual camera) labeled in the table as *with* on nine test videos.

	Fusion	# images	TP	FP	TN	FN	Accuracy	Recall	Precision	F1
vid1	without	123	36	231	6	62	12,54%	**36,73%**	13,48%	19,73%
	with	123	27	112	16	71	**19,03%**	27,55%	**19,42%**	**22,78%**
vid2	without	124	106	373	0	71	19.27%	**59.89%**	22.13%	32.32%
	with	124	99	104	5	78	**36,36%**	55,93%	**48,77%**	**52,11%**
vid3	without	124	105	333	0	33	22,29%	**76,09%**	23,97%	36,46%
	with	124	104	112	0	34	**41,60%**	75,36%	**48,15%**	**58,76%**
vid4	without	125	60	53	23	25	51,55%	70,59%	53,10%	60,61%
	with	125	60	46	32	25	**56,44%**	70,59%	**56,60%**	**62,83%**
vid5	without	110	80	19	25	4	82,03%	95,24%	80,81%	87,43%
	with	110	80	18	25	4	**82,68%**	95,24%	**81,63%**	**87,91%**
vid6	without	124	76	8	28	20	79,39%	80,00%	90,48%	84,92%
	with	124	76	6	28	20	**80,62%**	80,00%	**92,68%**	**85,88%**
vid7	without	260	262	35	21	41	78,83%	86,47%	88,22%	87,33%
	with	260	262	25	21	41	**81,09%**	86,47%	**91,29%**	**88,81%**
vid8	without	193	110	189	4	3	37,25%	**97,34%**	36,78%	53,39%
	with	193	101	12	84	12	**88,51%**	89,83%	**89,38%**	**89,38%**
vid9	without	372	346	144	55	29	69,86%	**92,26%**	70,61%	80,00%
	with	372	331	15	99	44	**87,93%**	88,26%	**95,66%**	**91,81%**



(a) (b)

Fig. 4. An illustration of the situation where the pistol is placed out of the area of interest. The result of the detection with (a) and without (b) fusion.

which could create the need for much longer training periods in comparison with the rest. In addition, the prediction process of NAS is slower, which makes it inappropriate for the problem studied in this work.

In addition, the performance of Faster R-CNN(VGG-16 + ImageNet) and Faster R-CNN(COCO based models) improved when using our image fusion technique. In fact, in the case of ResNet, the benefits of using our image fusion technique is more important due to a high percentage of recurrent false positives in the background. In all the experiments provided in next sub-sections, we used Faster R-CNN based on VGG-16 with transfer learning from ImageNet.

5.3. Impact of the proposed image fusion on the detection

Table 2 shows the number of TP, FP, TN, FN, accuracy, recall, precision and F1 measure with and without applying the proposed image fusion technique on nine test videos, from vid1 to vid9. In general, the detection with fusion reaches better accuracy, precision and F1 measure, on all the analyzed videos. In average, the accuracy has improved by 13, 47%, the precision by 16% and F1 by 10.89%. The number of false positives is reduced when applying fusion for all the test videos and this improvement increases when the number of input frames increases. In average, the number of false positives is reduced by 49.47%.

In particular, the improvements obtained with fusion on videos, vid5, vid6 and vid7 are modest due to the simplicity of the background in the office scenario. While the improvements obtained in vid1, vid2, vid3, vid8 and vid9 are more important due to the complex background in the scenario. The results of the most realistic video surveillance scenario, vid8 and vid9, are important.

The fusion approach substantially reduces the number of false positives in all videos and scenarios as it eliminates an important number of persistent false positives in some areas of the background.

On the other hand, applying fusion increased the number of false negatives especially in vid8 and vid9 due mainly to the imperfections in the disparity map calculation or when the pistol is located outside the area of interest. These imperfections can be addressed by using a more accurate disparity map algorithm. When the person moves out of the detection area behind the limit distance the system is unable to detect the pistol as it is illustrated in Fig. 4(a) and (b). The analysis of the detection of the non-fused images showed that when the handgun is detected too far from the camera, the detection becomes intermittent and low quality as it can be seen in Fig. 4(a). This situation could lead the detector to more false detections in real live environments, so the limit distance in the fused images must be kept in a range that provides confident detections.

The videos of the detection results with and without fusion on the nine test videos are available through this link (http://sci2s.ugr.es/weapons-detection).

5.4. Analysis of some examples

This subsection shows and analyzes the impact of the proposed fusion technique on the number of false positives using several examples.

5.4.1. Reducing false positives

Figs. 5 show a comparison of the detection results on several frames with, (b), (d), (f), (h) and without fusion, (a), (c), (e), (g). In general, one can observe from these examples that the used image fusion technique improves the overall detection results by generating a better candidate-regions, which helps the model to focus on the true area of interest. In

Fig. 5. Examples of the detection results with the proposed binocular image fusion, (b), (d), (f), (h) and without image fusion, (a), (c), (e), (g).

particular, as it can be seen in Figs. 5(a)–(f) the proposed fusion technique successfully eliminates all the false positives located in the eliminated background.

An ideal detection system must be able to differentiate between an image of a gun and a real gun. As it can be seen in Figs. 5(g)–(h), the proposed fusion technique can help the neural network to differentiate between a 2d printed gun and a real gun. In addition, from Figs. 6(a) and (b), one can observe that the proposed image fusion technique is able not only to eliminate the false positives located in the eliminated background but it also reduces the false positives located in the part of the background that was not blurred in the input frame due to its closeness to the area of interest.

5.4.2. *False negatives*

As it can be observed in Figs. 7(a)-(d) applying the proposed fusion technique and consequently producing more focused pre-selected areas of interest sometimes can eliminate false negatives produced by the handgun detection algorithm. Overall the proposed image fusion technique also improves the confidence of the detection model by increasing the probabilities of the detected true positives as it can be seen from these figures.

Fig. 6. Examples of the detection results with, (b), (d), (f), and without the proposed binocular image fusion technique, (a), (c), (e).

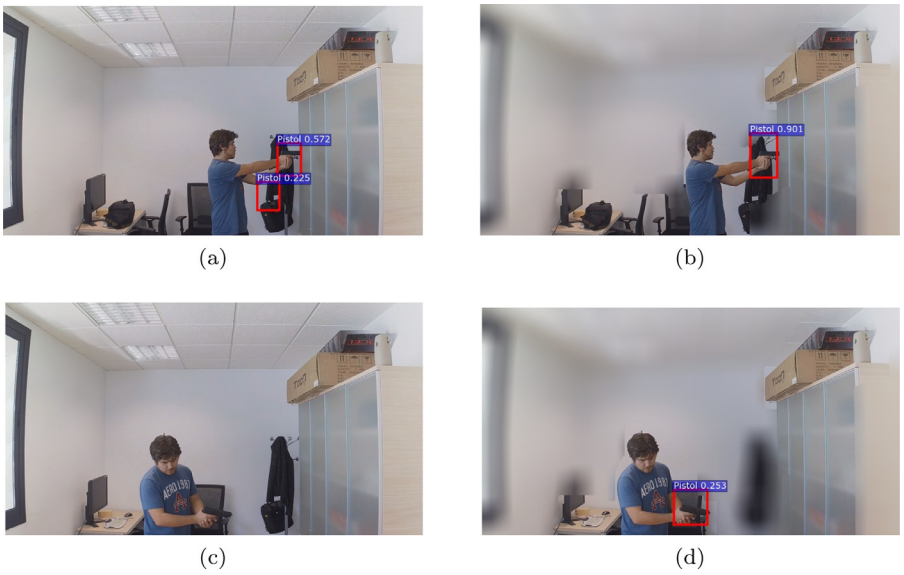


Fig. 7. Examples of the detection results with, (b), (d) and without, (a), (c) image fusion.

In summary, our results show that use of symmetric cameras helps reducing the number of false positives thanks to the similarities it presents with the human vision and the way we perceive the deepness of the environment. Our approach can be easily extended to general object detection since modern smartphones are increasingly including dual cameras.

6. Conclusions

This work presented a new image fusion technique that improves the results of handgun detection in surveillance videos. We built a symmetric dual camera system to obtain the three-dimensional information and used this information to remove the background from the scene. We then

apply a series of filters to obtain the map of the main area of interest. We blur, in the input frames, the areas that do not belong to the area of interest. This strategy substantially helps the model to focus on the real area of interest where the probability of finding a handgun is higher. This approach reduced considerably the number of false positives and improves the reliability of the detection in security field.

As future work, we will evaluate the use of different dual cameras setup. We will also evaluate different combinations of infrared images, visible light images and the motion information to pre-select the areas of interest.

Acknowledgements

This work was partially supported by the Ministerio de Ciencia y Tecnologia under the project: TIN2017-89517-P. Siham Tabik was supported by the Ramon y Cajal Programme (RYC-2015-18136).

References

[1] J. Dai, Y. Li, K. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks, in: Advances in neural information processing systems, 2016, pp. 379–387.
[2] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[3] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
[4] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[5] H. Hirschmuller, Stereo processing by semiglobal matching and mutual information, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 328–341.
[6] 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: IEEE CVPR, Vol. 4, 2017.
[7] Y.J. Jung, Enhancement of low light level images using color-plus-mono dual camera, Opt. Express 25 (10) (2017) 12029–12051.
[8] D.N. Kim, Y.S. Chae, M.Y. Kim, X-Ray and optical stereo-based 3d sensor fusion system for image-guided neurosurgery, Int. J. Comput. Assist. Radiol. Surg. 11 (4) (2016) 529–541.
[9] H. Kim, J. Jo, J. Jang, S. Park, J. Paik, Seamless registration of dual camera images using optimal mask-based image fusion, in: Consumer Electronics-Asia (ICCE-Asia), IEEE International Conference on, IEEE, 2016, pp. 1–2.
[10] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.
[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37.
[13] B. Moon, S. Yu, S. Ko, S. Park, J. Paik, Continuous digital zooming using local self-similarity-based super-resolution for an asymmetric dual camera system, JOSA A 34 (6) (2017) 991–1003.
[14] R. Olmos, S. Tabik, F. Herrera, Automatic handgun detection alarm in videos using deep learning, Neurocomputing 275 (2018) 66–72.
[15] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[16] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.
[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vision (IJCV) 115 (3) (2015) 211–252, doi:10.1007/s11263-015-0816-y.
[18] H. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R.M. Summers, Deep convolutional neural networks for computer-aided detection: cnn architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (5) (2016) 1285–1298.
[19] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556 (2014).
[20] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning., in: AAAI, 4, 2017, p. 12.
[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[22] J.R. Uijlings, K.E. Van De Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, Int. J. Comput. Vis. 104 (2) (2013) 154–171.
[23] S. Yu, B. Moon, D. Kim, S. Kim, W. Choe, S. Lee, J. Paik, Continuous digital zooming of asymmetric dual camera images using registration and variational image restoration, Multidimens. Syst. Signal Process. 29 (4) (2017) 1–29.
[24] Z. Zhang, W. Tao, K. Sun, W. Hu, L. Yao, Pedestrian detection aided by fusion of binocular information, Pattern Recognit. 60 (2016) 227–238.
[25] B. Zoph, V. Vasudevan, J. Shlens, and Q.V. Le. Learning transferable architectures for scalable image recognition. (2017) arXiv preprint arXiv:1707.07012, 2(6).