Public Security Video and Image Analysis Challenge: A Retrospective

Gengjian Xue(✉), Wenfei Wang, Jie Shao, Chen Liang, Jinjing Wu, Hui Yang, Xiaoteng Zhang, Lin Mei, and Chuanping Hu

The Third Research Institute of the Ministry of Public Security, Shanghai 200031, China
xgjsword@163.com, wolfeiwang@139.com, jieshao.mail@gmail.com, i_liang_chen@foxmail.com, xyzl.xt@163.com, dongjiyinxin@163.com, zxt881108@126.com, gasswlw@126.com, cphu@vip.sina.com

Abstract. The Public Security Video and Image Analysis Challenge (PSVIAC) is a benchmark in object detection and instance search on public security surveillance videos. This challenge is first held in 2016, attracting participation from more than twenty institutions. This paper provides a review of this challenge, including tasks definition, datasets creation, ground truth annotation, and results comparison and analysis. We conclude the paper with some future improvements.

1 Introduction

Recently, many kinds of challenges for video and image analysis have been paid much attention [1–4], because they can provide large mount of data for specific tasks and build a platform for testing various algorithms fairly and publicly. For example, the PASCAL Visual Object Classes Challenge [1] has been an annual event sine 2006, which focuses on visual object classification, detection, segmentation, *etc.*; the ImageNet Large Scale Visual Recognition Challenge [2] has been run annually from 2010 to present, which can be considered doing objects recognition on large scale datasets; the TREC Video Retrieval Evaluation [3] aims to improve the content-based analysis and retrieval technologies. However, the datasets they used are often obtained in specific situations, whose original meaning is to model real world situations. As a result, researchers have got good performance on some tasks, but their algorithms may not work well in practice. Perhaps one of the main reasons is that the data is not obtained from real scenes.

To better deal with visual recognition problems in public security areas, the Public Security Video and Image Analysis Challenge (PSVIAC) was held for the first time. It was also as a special contest on the 2016 Symposium on Research and Application in Computer Vision. The data used in PSVIAC was obtained from real public security scenes, so this challenge could be considered as a benchmark in object detection and instance search for public security applications.

This paper is organized as follows: we start with a review of this challenge in Sect. 2, describing in brief tasks definition, datasets creation, annotation procedure, and evaluation measures. Section 3 provides an overview of the results. We

C.-S. Chen et al. (Eds.): ACCV 2016 Workshops, Part III, LNCS 10118, pp. 441–452, 2017.
DOI: 10.1007/978-3-319-54526-4_32

then use these results for several additional analysis. Section 4 discusses some suggestions that may be useful for future challenges. Finally, we conclude this paper in Sect. 5.

2 Challenge Review

This challenge consists of two components: (1) an available dataset of images for training and test, and ground truth annotations for training; (2) a workshop for summarizing the challenge and discussing the results. This section describes in detail the tasks, datasets, annotations, and evaluation procedures.

2.1 Challenge Tasks

There are two principal challenges: (1) object detection – "does the image contain any instances of a particular object class and where are the instances of a particular object class"; (2) instance search – "given one example of the specific target, it is to find out more images that contain this target".

Object Detection. Three are three object classes for detection: **Non_vehicle**, **Vehicle**, and **Pedestrian**. For each of the three classes, participants are required to find each object of that class in a given test image (if any) and predict the bounding box of that object with associated real-valued confidence. In this task, participants may have to localize multiple object classes in the image, which makes the task more demanding. Any annotation provided in the PSVIAC training data could be used. Participants are not permitted to perform additional manual annotation of either training or test data.

Instance Search. Two classes of instances are used for this task: **Non_vehicle**, **Vehicle**. Given an instance target, participants are required to find out the images that most likely contain this instance and predict the bounding box of the instance. For each query instance, at most 100 candidate results are allowed to be submitted, arranging in descending order according to their possibilities. Thus, each result includes such information: the image name, the predicted bounding box coordinates, and its sorted number.

In this contest, the additional requirement to locate the instance in an image makes the task more challenging, since guessing the right answer is far more difficult to achieve. However, it is really needed in police practical applications.

2.2 Datasets Construction

For the purpose of challenge, the dataset is divided into two subsets: object detection dataset, and instance search dataset. In order to reduce the amount of calculation and ensure the data quality, we extract I-frames from collecting videos to construct the dataset.

Object Detection Dataset. In practical applications, **Non_vehicle**, **Vehicle**, and **Pedestrian** are the three common types. However, since real scenes are often complicated, some aspects should be considered for creating a valuable dataset.

1. *Weather Condition*: The dataset should contain images under various weather conditions, such as sunny, cloudy, and rainy.
2. *Scale Condition*: The dataset should contain images that are taken under a variety of distance, including long distance, middle distance, and close distance.
3. *Angle Condition*: The dataset should contain images that are taken from multiple views.
4. *Multiple Objects*: The dataset should contain images that multiple objects exist in an image.
5. *Occlusion Condition*: The dataset should contain images that contain occluded objects, including unrecognizable and recognizable objects.

Based on these requirements, we collect a large mount of candidate images and select high quality images from them to construct this dataset. Since too many or few objects in an image may be not useful for training and test, the high quality image should satisfy these conditions at the same time:

(a). The number of valid objects in an image is between 3 and 12.
(b). The number of invalid objects in an image is less than 10.
(c). The number of valid occluded objects in an image is less than 6.

The valid object is defined as the area of the object is bigger than 900 pixels and can be recognized by human eyes. The valid occluded object is defined as the ratio of the occluded area to its total area is smaller than 0.5, and the object can be recognized by human eyes.

Some examples of the object detection dataset are shown as follows. Figure 1 shows the examples in this dataset with various weather conditions. Figure 2 shows the examples in this dataset with various scale conditions. Figure 3 shows the examples in this dataset with various angle conditions. Figure 4 shows the examples in this dataset with multiple objects. Figure 5 shows the examples in this dataset with various occlusion conditions.

In this object detection dataset, there are total 39151 images. We first select 20000 images for training. The remaining 19151 images are for candidate test, then from which 10000 images are selected for formal object detection test.

Fig. 1. Examples of the object detection dataset with various weather conditions.

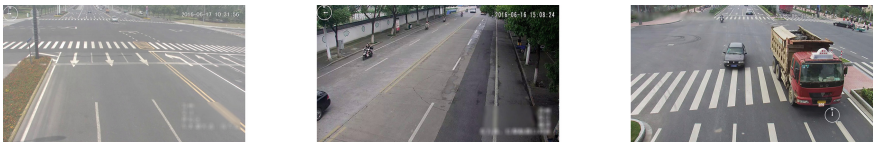

Fig. 2. Examples of the object detection dataset with various scale conditions.

Fig. 3. Examples of the object detection dataset with various angle conditions.

Fig. 4. Examples of the object detection dataset with multiple objects.

Fig. 5. Examples of the object detection dataset with various occlusion conditions.

Instance Search Dataset. For the instance search dataset creation, we have noticed some aspects in practical applications.

1. *Weather Condition*: The instance may appear in various weather conditions.
2. *Scale Condition*: The images containing the instance should be captured under a variety of distance, including long distance, middle distance, and close distance.
3. *Angle Condition*: The images containing the instance should be taken from multiple angles.
4. *Occlusion Condition*: The instance in the image may be occluded by other objects.

Based on these considerations, we collect a large mount of high quality images to construct the instance search dataset. Some example of this dataset are shown as follows. Figure 6 shows the examples in this dataset with various weather conditions. Figure 7 shows the examples in this dataset with various scale conditions.

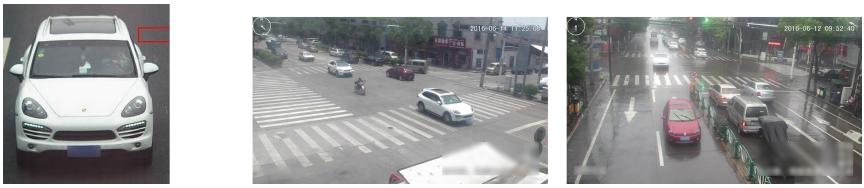

Fig. 6. Examples of the instance search dataset with various weather conditions. The left image is the given instance. The middle and right images are the images containing this instance.

Fig. 7. Examples of the instance search dataset with various scale conditions. The left image is the given instance. The middle and right images are the images containing this instance.

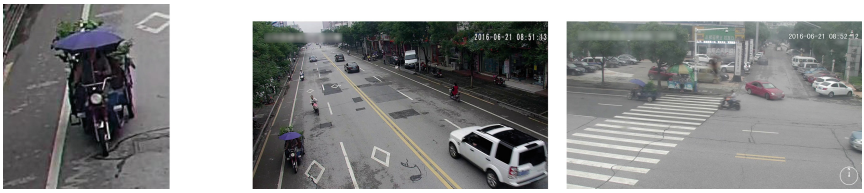

Fig. 8. Examples of the instance search dataset with various angle conditions. The left image is the given instance. The middle and right images are the images containing this instance.

Fig. 9. Examples of the instance search dataset with various occlusion conditions. The left image is the given instance. The middle and right images are the images containing this instance.

Figure 8 shows the examples in this dataset with various angle conditions. Figure 9 shows the examples in this dataset with various occlusion conditions.

By statistics, the instance search dataset contains 47458 images, where the total number of images for **Vehicle** instance search is 24396, and the total num-

ber of images for **Non_vehicle** instance search is 23062. 100 candidate instances have been created, including 60 **Vehicle** instances and 40 **Non_vehicle** instances, then from which 10 **Vehicle** instances and 5 **Non_vehicle** instances are selected for formal instance search test.

2.3 Annotation Procedure

The annotation procedure consists of two steps:

(1) Sensitive information annotation. The sensitive information in an image includes: (a) The information from which we can determine the places that the image are taken, such as road names, place names. (b) The information from which we can determine the specific object, such as plate numbers, advertising messages. We annotate the bounding boxes that contain this information. Then, these areas are blurred by convolution.

(2) Valid objects and valid occluded objects annotation. For the object detection dataset, we annotate all the valid objects and valid occluded objects in each image. For the instance search dataset, we just annotate the object corresponding the same instance in an image. The ground truth area should be the smallest bounding box including the object. For occluded objects, we predict the occluded area, and label the ground truth area. The annotated deviation is within 4 pixels.

In the object detection dataset, statistics indicates that in the training dataset, the total number of **Vehicle**, **Non_vehicle**, and **Pedestrian** are 75653, 29725, and 9632 respectively. While in the formal test dataset, the total number of **Vehicle**, **Non_vehicle**, and **Pedestrian** are 42129, 11689, and 2767 respectively. In the instance search dataset, for each **Vehicle** instance query, the number of images containing this instance is about 20; and for each **Non_vehicle** instance query, the number of images containing this instance is about 15.

2.4 Evaluation Measures

Object Detection Evaluation. The criteria for objects detection is designed to penalize the algorithm for missing object instances, for duplicate detections of one instance, and for false positive detections.

Detections are assigned to groundtruth objects and judged to be true or false positives by measuring bounding box overlap. Let $IOU(B_p, B_{gt})$ be the overlap area between the predicted bounding box B_p and ground truth bounding box B_{gt}, it is computed as:

$$IOU(B_p, B_{gt}) = \frac{B_p \cap B_{gt}}{B_p \cup B_{gt}} \tag{1}$$

where $B_p \cap B_{gt}$ means the intersection area of the predicted and ground truth bounding boxes, and $B_p \cup B_{gt}$ denotes their union. A detection is considered as correct when its $IOU(B_p, B_{gt})$ value exceeds a given threshold T_{det}, where it is set to be 0.5 in this contest.

For each object class, we first compute its *precision-recall* curve from a method's rank output. The *precision* is defined as the fraction of correct detections out of the total detections returned by the algorithm, and the *recall* is defined as the fraction of the correct detections out of the total ground truth instances in the dataset. The interpolated average precision (denoted as AP_{det}) [5] is adopted as the average measure over one detection class, which summaries the shape of the precision-recall curve, and is defined as the mean precision at a set of uniformly-spaced *recall* value [0, 0.1, 0.2, ..., 1]:

$$AP_{det} = \frac{1}{11} \Sigma_{r \in 0,0.1,\ldots,1} P_{interp}(r) \tag{2}$$

The precision at each *recall* level r is interpolated by taking the maximum *precision* measured for a method for which the corresponding *recall* exceeds r:

$$P_{interp}(r) = \max_{r':r' \geq r} p(r') \tag{3}$$

where $p(r')$ is the measured *precision* at *recall* r'.

The overall performance on the object detection task is got by averaging the AP_{det} values on three object classes.

Instance Search Evaluation. For a given instance, at most 100 candidate results are allowed to be returned. One search result considered to be correct should satisfy these conditions: (1) The returned image name can be found in the ground truth image name list; (2) The overlap area between its predicted bounding box and the ground truth bounding box exceeds a given threshold T_{ins}, which is set to 0.1 in this contest, where the computation of the overlap area is according to Eq. (1).

The performance over one instance search is measured by computing the average precision in retrieval (denoted as AP_{ins}) [6,7], which is defined as follows:

$$AP_{ins} = \frac{1}{R} \Sigma_{j=1}^{n} I_j \times \frac{R_j}{j} \tag{4}$$

where R means the total number of ground truth images containing the specific instance, n stands for the total number of returned images by the algorithm (We set $n = 100$ in this contest), j is the index number, I_j is 1 when the j_{th} result is correct, otherwise I_j equals to 0, R_j means the total number of correct results in the first j results.

The overall performance on the instance search task is got by averaging the AP_{ins} values on all test instances.

Score and Ranking. In this challenge, the organization committee decides to give one final ranking according to both the object detection and instance search contests. Since different evaluation measures have been used for these two competitions, directly combining or adding the performance values is not feasible. To deal with this problem, we adopt the competition score instead of

the performance value. The competition score of the individual team in each contest is computed according to its ranking order which is sorted based on its performance.

For each contest, all teams are first sorted according the performance in descending order. With a ranked number, the team's score in this contest is computed as:

$$DET_{score} = 25 * (2 - log(D_{th})) \tag{5}$$

$$INS_{score} = 25 * (2 - log(I_{th})) \tag{6}$$

where D_{th} and I_{th} are its ranked number in the object detection and instance search contests respectively, DET_{score} and INS_{score} denote the team's scores got in the object detection and instance search contests respectively.

The final score of this team in this challenge is computed by adding the above two scores:

$$TOTAL_{score} = INS_{score} + DET_{score} \tag{7}$$

where $TOTAL_{score}$ is the total score of this team in the challenge.

Finally, all teams are sorted according to their scores in descending order, and the winner of this challenge is the team that has the highest score.

We should mention that this is an empirical calculation method. According to this method, each team is encouraged to participate both two contests and submit their results. If it just submits one result, then the score of this team in the other contest is considered to be 0, which would have great impact on its final score.

2.5 Submission

Each team is allowed to submit one final result in each contest. Two contests are both just for automatic runs, not including interactive runs.

For data safety, we have provided virtual environments for operation. Each team had to sign a confidentiality agreement before using the data. The dataset was released in the virtual environment machine which was not permitted for downloading. But participants were allowed to upload their code and other sources to the virtual environment. Three operation systems have been provided for choosing: CentOS, Ubuntu, and Win 7. For each virtual environment, a GPU device has been provided as well. After inquiry and collection, 20 teams, 3 teams, and 5 teams have chosen Ubuntu, CentOS, and Win 7, respectively.

The ground truth training data and the submitted results are both saved as text file according to the required format. The test data was available for half a month days before final submission. During this period, we have encouraged teams to submit their periodical results for evaluation, which may help them to improve algorithms.

3 Results and Analysis

3.1 Results

To the end, only 13 teams have finished this challenge and submitted their results. Table 1 shows their performance on each object class, the overall performance, and rankings in the object detection contest. Table 2 shows their overall performance and rankings in the instance search contest. Table 3 shows their scores and final rankings in this challenge.

It can been that the TH-MIG team has got the best performance in the object detection contest, the HawkEye team and the ZJU teams have got the second and the third places, respectively. The overall performance values of these four teams were all higher than 0.75. In the instance search contest, the overall performance values were relatively low. However, the DongGua team has achieved outstanding performance. The ZJU and SkyWalker teams were the second and third respectively.

Overall, the DongGua and TH-MIG teams both got the highest score 84.95, and they tied for the first. The ZJU and HawkEye teams were the third and fourth. These four teams have been invited to attend the workshop and make a speech.

Table 1. The performance and ranking in the object detection contest.

Team name	Vehicle	Non_vehicle	Pedestrian	AP_{det}	Ranking
KAOYU	69.44	63.08	28.56	53.69	7
AHU_CVPR	86.17	75.63	53.22	71.67	5
DongGua	85.48	74.94	60.27	73.56	4
BaiPao	78.60	38.29	14.49	43.79	10
ZJU	86.56	82.17	60.57	76.43	3
Primary_CvVer	50.20	39.02	10.72	33.31	12
HawkEye	86.91	81.72	62.07	76.90	2
TeamAdelaide	84.37	71.62	32.54	62.85	6
TH-MIG	88.11	84.41	65.77	79.43	1
SkyWalker	68.46	48.92	28.45	48.61	8
Endless	0.29	0.01	0.00	0.10	13
FTD	60.45	48.54	22.87	43.95	9
HuanJing	64.11	44.66	12.06	40.27	11

3.2 Analysis

In this subsection, some analysis on the results will be given. Table 4 shows the average performance of each individual contest over all teams. We can seen that participants have got the best results in doing the **Vehicle** detections,

Table 2. The performance and ranking in the instance search contest.

Team name	AP_{ins}	Ranking
KAOYU	0.89	9
AHU_CVPR	0.00	10
DongGua	45.79	1
BaiPao	0.00	10
ZJU	6.07	2
Primary_CvVer	0.00	10
HawkEye	1.46	8
TeamAdelaide	4.00	5
TH-MIG	4.22	4
SkyWalker	5.20	3
Endless	0.00	10
FTD	2.01	6
HuanJing	1.69	7

Table 3. The final scores and rankings in this challenge.

Team name	DET_{score}	INS_{score}	$TOTAL_{score}$	Final ranking
KAOYU	28.87	26.14	55.02	9
AHU_CVPR	32.53	25.00	57.53	7
DongGua	34.95	50.00	84.95	1
BaiPao	25.00	25.00	50.00	11
ZJU	38.07	42.47	80.55	3
Primary_CvVer	23.02	25.00	48.02	12
HawkEye	42.47	27.42	69.90	4
TeamAdelaide	30.55	32.53	63.08	6
TH-MIG	50.00	34.95	84.95	1
SkyWalker	27.42	38.07	65.49	5
Endless	22.15	25.00	47.15	13
FTD	26.14	30.55	56.69	8
HuanJing	23.97	28.87	52.84	10

and better results have been obtained in the **Non_vehicle** detections. The **Pedestrian** detection seems to be the most difficult detection task. The reason may be as follows. First, the **Vehicle** detection is a common task. Much data could be used and researchers have studied it for years. Second, **Vehicle** is a rigid object. Although our scenes are complicated, its shape keeps relatively unchanged. Third, compared with other two classes, the area of **Vehicle**

is bigger, which is easier to be detected. For the **Non_vehicle** detection, fewer publicly available data can be used. Furthermore, it is not absolutely a rigid object. Although many **Pedestrian** datasets have been released for training, many public algorithms may be not work well in our cases. The pedestrians in our dataset are often relatively small, with multiple views, and affected by various factors. Whereas most public datasets were obtained under ideal conditions.

For the instance search task, it is more demanding. It should mention that only two teams are actually well beyond the average value. One reason is that only one example per instance has been provided, so that the training process is challenging. Another reason is that the instance may appear in various forms, such as multi-scale variation, multi-view variation, and occlusion. The DongGua team has got the outstanding result by first using deep learning technologies [8] and other data to train many models off-line, and then fusing these models on our datasets. However, the performance may be further improved if small sample learning problems would be well resolved.

Table 4. The average performance of each individual contest

Name	Vehicle	Non_vehicle	Pedestrian	Instance search
Average performance	69.93	57.92	34.74	5.49

4 Discussions

This section discusses some topics for future improvement of the challenge.

The first topic is about the dataset augmentation. In this first challenge, we have totally released 86609 images for training and test. However, the mount of data seemed to be not big enough to support models training. Another problem is that the number of objects belonging to different classes is not uniform. For example, the number of pedestrian objects is much less than that of other two classes, which may affect the algorithm's performance. So it is necessary to effectively expand the dataset.

The second topic is about the evaluation measures. In this challenge, we have adopted an empirical approach to calculate scores and rankings. However, there may exist more reasonable methods. Besides, other objective evaluation measures could be adopted for evaluating algorithms from multiple aspects.

The third topic is that more in-depth analysis on the results and algorithms should be introduced, such as algorithms comparison, distribution analysis of results, and inter-class comparisons.

5 Conclusions

The PSVIAC has contributed to the development of video and image analysis technologies in public security. More than twenty institutes from home and

abroad have participated this competition, and many effective methods have been proposed. We believe that this first challenge is a good start and it will be getting more and more better with continuous improvements.

Acknowledgement. Our research was sponsored by following projects: Program of Science and Technology Commission of Shanghai Municipality (No. 15530701300, 15XD1520200, 14DZ2252900); 2012 IoT Program of Ministry of Industry and Information Technology of China; Key Project of the Ministry of Public Security (No. 2014JSYJA007); Shanghai Science and Technology Innovation Action Plan (No. 16511101700).

References

1. Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., et al.: The pascal, visual object classes challenge: a retrospective. Int. J. Comput. Vis. **111**, 98–136 (2015)
2. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**, 211–252 (2015)
3. Over, P., Fiscus, J., Sanders, G., et al.: Trecvid 2014-an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID (2014)
4. Patino, L., Ferryman, J.: Pets: dataset and challenge. In: IEEE International Conference on Advanced Video and Signal Based Surveillance 2014, pp. 355–360 (2014)
5. Salton, G., Mcgill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill Inc., New York (1986)
6. Zhu, M.: Recall, precision, and average precision. Department of Statistics and Actuarial Science, University of Waterloo (2004)
7. Turpin, A., Scholer, F.: User performance versus precision measures for simple search tasks. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 11–18 (2006)
8. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798–1828 (2013)