Extremely Overlapping Vehicle Counting

Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Oñoro-Rubio

GRAM, University of Alcalá, Spain.

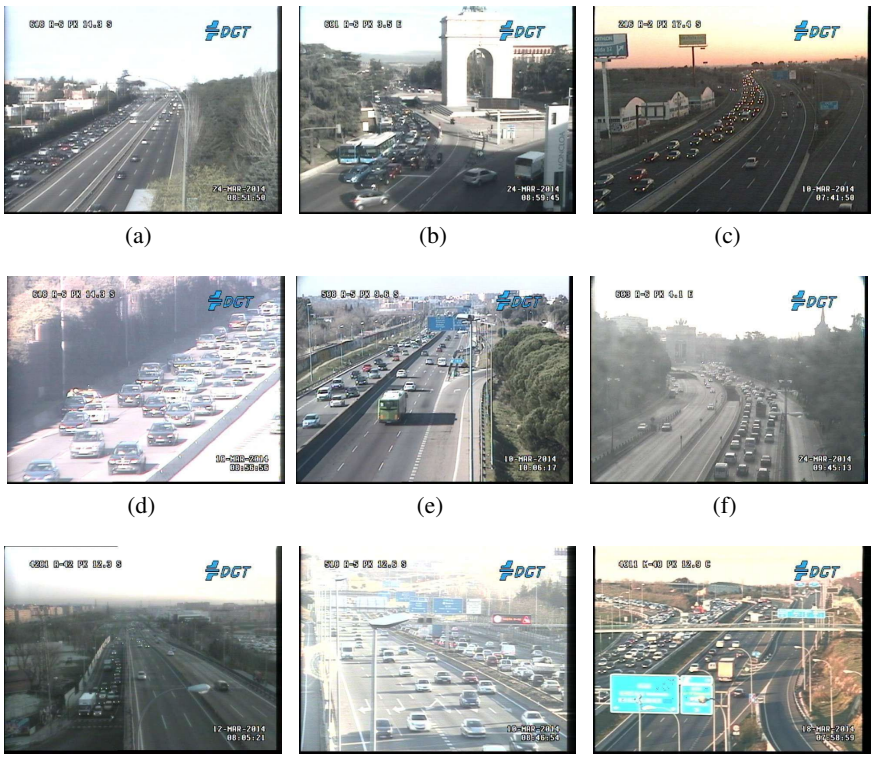
Abstract. The challenging problem that we explore in this paper is to precisely estimate the number of vehicles in an image of a traffic congestion situation. We start introducing TRANCOS, a novel database for extremely overlapping vehicle counting. It provides more than 1200 images where the number of vehicles and their locations have been annotated. We establish a clear experimental setup which will let others evaluate their own vehicle counting approaches. We also propose a novel evaluation metric, the Grid Average Mean absolute Error (GAME), which overcomes the limitations of previously proposed metrics for object counting. Finally, we perform an experimental validation, using the proposed TRANCOS dataset, for two types of vehicle counting strategies: counting by detection; and counting by regression. Our results show that counting by regression strategies are more precise localizing and estimating the number of vehicles. The TRANCOS database and the source code for reproducing the results are available at `http://agamenon.tsc.uah.es/Personales/rlopez/data/trancos`.

1 Introduction

Extremely overlapping vehicle counting is a very challenging problem. See Figure 1, where we show different traffic congestion images. To precisely estimate the number of vehicles present in this type of scenes is not an easy task, even for a human. Building automatic counting solutions able to deal with this problem would allow the development of systems that precisely monitor the evolution of a traffic jam. This information would result invaluable for the public authorities in charge of the maintenance and planning of road infrastructures.

To the best of our knowledge, this work presents the first attempt to provide a dataset, and an associated benchmark, to experimentally evaluate the performance of different approaches dealing with the problem of extremely overlapping vehicle counting. There are other datasets which have been previously used to evaluate the precision of vehicle detection approaches, e.g. [1,2,3,4], but none of them contain traffic congestion images where the overlap between the vehicles is considerable.

For instance, the car detection task in the PASCAL VOC [1] benchmark has reached an Average Precision above 50%. Most of the winner methods in this dataset are based on detectors using Histogram of Oriented Gradients (HOG) features [5] and sliding window strategies. Figure 2 (a) depicts the results of a HOG based detector [6] in one of our images. It seems clear that novel solutions need to be explored. We show in this paper that this type of strategy, named counting by detection, does not improve the precision reported by counting by regression techniques, e.g. [7,8].



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 1. TRaffic ANd COngestionS (TRANCOS) database images. These pictures show how challenging the proposed problem of extremely overlapping vehicle counting is, even for a human.

But also some additional problems have to be considered within this novel context. Typically, public authorities deploy a big net of cameras for the traffic surveillance. This fact implies that the network bandwidth must be controlled, hence the images delivered have a low resolution. This aspect directly affects the performance of gradient based features object detectors (e.g. [5]). Furthermore, in these images the vehicles normally occupy a few pixels and the context becomes fundamental to actually recognize the objects. See Figure 2 (b), where a car can be easily confused with a black couch. In summary, the proposed scenario presents an interesting object counting problem which has not been previously explored.

The key contributions of our work can be summarized as follows. First, we release the novel TRaffic ANd COngestionS (TRANCOS) dataset, which has been specifically designed to evaluate the performance of extremely overlapping vehicle counting solutions. With more than 1200 traffic jam images and 46700 annotated vehicles, TRANCOS comes with the clear experimental setup detailed in this paper, in order to set a new benchmark. Second, we propose a novel evaluation metric, the Grid Average Mean absolute Error (GAME) metric. We show that the GAME metric overcomes some of

(a) (b)

Fig. 2. (a) Results of the HOG detector [6] in a traffic congestion image. (b) In these images the vehicles typically occupy a few pixels, and it becomes difficult to identify the typical cues that we use for finding cars (e.g. wheels, license plates). Instead, note that the context results fundamental to determine whether a blob represents a car or a black couch, for example.

the limitations of previously proposed metrics for object counting, such as the Mean Absolute Error (MAE) [9]. And third, we complete the paper with an experimental validation, using the TRANCOS dataset, for three state-of-the-art methods for object counting: a counting by detection approach based on the HOG detector described in [6]; and the two counting by regression approaches [7] and [8], which have previously reported the best results for the problem of crowd counting. The experimental results confirm that the counting by regression strategies are more precise localizing and estimating the count of the vehicles. Our aim with this work is to offer to the computer vision community a novel benchmark for the problem of extremely overlapping vehicle counting.

The rest of the paper is organized as follows. Section 2 reports on related work. Section 3 includes a detailed description of the TRANCOS dataset and the proposed GAME metric. In Section 4 we show the experimental validation, and Section 5 concludes the paper.

2 Related work

To the best of our knowledge, the vehicle counting problem in traffic congestion images has not been previously systematically studied. There are works for vehicle detection in conditions of relatively high traffic in highways, such as the models in [10,11,12]. But the problem proposed in this paper is completely different. First, all these previous works offer solutions adapted to video, and we simply provide still (low resolution) images, obtained from real traffic surveillance cameras. They all incorporate a background subtraction step to their system pipelines, but with the TRANCOS dataset this is not possible. Moreover, our images cover a variety of viewpoints and scenarios, and not simply two or three fixed scenes. This implies that the parameterization of the scene cannot be considered. Finally, the grade of overlap between the vehicles that the TRANCOS dataset offers is considerable, in sharp contrast to the rest of datasets.

We understand the extremely overlapping vehicle counting problem as a variant of the problem of crowd counting [7,8,9,13,14,15]. In this paper we consider two families of solutions. First, we have the popular counting by detection approaches, which count instances of objects through scanning the image space using a detector trained with local

image features (e.g. [5]). Second, we have the counting by regression solutions (e.g. [7]), which count objects by learning a direct mapping from low-level imagery features to objects density. The approach in [8] follows this idea too, but, instead of a linear transformation, the model uses a Regression Forest combined with a spatial average of the estimated densities to make smoother predictions. In this paper we evaluate the performance of all these state-of-the-art approaches using the novel TRANCOS dataset.

3 TRANCOS Dataset

We introduce in this section the TRANCOS dataset. Although there are several datasets for assessing the performance of vehicle detection approaches (e.g. [1,2,3,4]), TRANCOS is the first one focused on the problem of vehicle counting in traffic jam images, captured using *real* traffic surveillance cameras. Figure 1 shows a sample of the images provided by TRANCOS, which illustrates how challenging the proposed problem is. Note that all the collected images contain traffic congestions, covering a variety of different scenes and viewpoints, with changes in the lighting conditions, and considerable different levels of overlap and crowdedness, even within the same image.

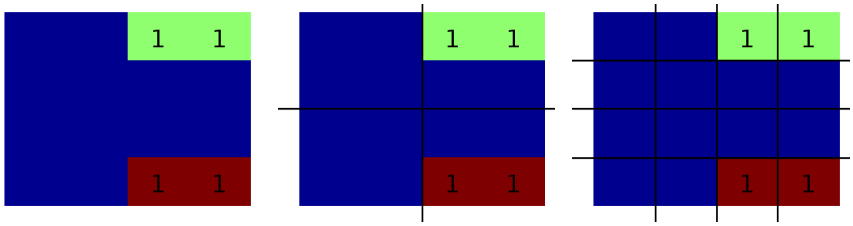
Specifically, the database consists of 1244 images. They have been acquired from a selection of public traffic surveillance cameras provided by the Directorate General of Traffic (DGT) of the Government of Spain. The cameras selected monitor different highways located in the area of Madrid, which typically present heavy traffic congestions.

Each image has been manually annotated. For this purpose, we follow a dotting annotation strategy, as in [7], and provide for each image the exact number of vehicles and their locations. In total, 46796 vehicles have been annotated. A Region of Interest (ROI) to identify the road region, is also provided for each image.

The main goal of TRANCOS is to evaluate vehicle counting approaches, especially under extremely overlapping conditions. So, any method using this dataset has to predict, for each test image, the number of vehicles, and their locations in the images. We propose the following experimental setup which has to be followed by any method using the dataset. The acquisition of the images has been done during three different weeks, which lets us distribute the images in three separate sets: training (403 images), validation (420) and test (421). We define two types of training strategies: 1) methods which are trained using only the provided training and validation data; 2) methods built using any data except the provided test data. In both cases, the test set must be used strictly for reporting of results alone - it must not be used in any way to train or tune systems, for example by running multiple parameter choices and reporting the best results obtained. This has to be done using the validation images, for instance.

With TRANCOS[1] we provide a set of tools for accessing the dataset and annotations described. For the evaluation metric, we introduce the novel GAME metric, which is detailed in Section 3.1.

[1] `http://agamenon.tsc.uah.es/Personales/rlopez/data/trancos`



(a) MAE = 0 (b) GAME(1) = 4 (c) GAME(2) = 4

Fig. 3. Toy example for the GAME and MAE metrics. We see in green the estimation and in red the ground truth, representing the count of the vehicles and their location in the image. The MAE in (a) is of 0, even when the objects have not been correctly located. In (b) and (c) we show how the GAME is able to penalize the count when the localization is wrong.

3.1 The GAME metric

In datasets for crowd counting, such as the UCSD Pedestrian Dataset [9], the metric chosen to evaluate the performance is the Mean Absolute Error (MAE), which is defined as follows,

$$\text{MAE} = \frac{1}{N} \cdot \sum_{n=1}^{N} |e_n - gt_n| \ , \tag{1}$$

where e_n corresponds to the estimated objects count for image n, and gt_n is the ground truth provided for image n, being N the total number of images considered.

While this metric seems fair for establishing a comparative, we have observed in our experiments that it often leads to mask mistaken estimations. The reason is that the MAE does not take into account *where* the estimations have been done in the images.

In order to provide a more accurate evaluation, we introduce here the Grid Average Mean absolute Error (GAME) metric. Our objective is clear: to offer an evaluation metric which simultaneously considers the object count, and the location estimated for the objects. With the GAME metric, we proceed to subdivide the image in 4^L non-overlapping regions, and compute the MAE in each of these subregions. We formulate the GAME as follows,

$$\text{GAME}(L) = \frac{1}{N} \cdot \sum_{n=1}^{N} (\sum_{l=1}^{4^L} \left| e_n^l - gt_n^l \right|) \, , \tag{2}$$

where, e_n^l is the estimated count in a region l of image n, and gt_n^l is the ground truth for the same region in the same image. The higher L, the more restrictive the GAME metric will be. Note that the MAE can be obtained as a particularization of the GAME when $L = 0$.

As it can be seen in Figure 3, our GAME metric is able to penalize those predictions with a good MAE but a wrong localization of the objects.

4 Results

4.1 Experimental Setup

Our main objective is to establish a novel benchmark for the problem of vehicle counting in traffic congestion situations. For doing so, we offer here the experimental validation for three different state-of-the-art methods using the described TRANCOS dataset.

The first method we implement is a counting by detection approach using the de facto standard detector based on HOG features [5]. Explicitly, we build our approach using the implementation of [6]. We design two different strategies to train this detector. The first one (HOG-1), consists in collecting positive and negative examples using the PASCAL VOC 2007 Dataset [1]. This HOG-1 lets us evaluate how challenging are the images provided in the novel TRANCOS dataset, when a well known dataset is used for learning the vehicle counting solutions.

Our second approach, HOG-2, uses the training data provided with the TRANCOS dataset, but the negatives are extracted using the PASCAL VOC 2007 and the GRAM-RTM [2] datasets. Note that with the TRANCOS dataset we do not provide bounding boxes annotations, but dots. Therefore, to collect positive examples for training the HOG-2, we proceed as follows. For each training image, we apply the detector HOG-1 at multiple scales and collect detections which later are manually filtered to identify those that contain correct positive examples. These positives are used to train the HOG-2 detector. This way, we are able to train the detector using data generated from the same distribution provided with the TRANCOS dataset.

We also analyze the performance of two counting by regression models, using only the TRANCOS data. First, we follow the approach in [7] to learn a linear regression model to predict vehicle densities in the images. For the visual features, we compute dense SIFT [16] descriptors, which are then quantified using a K-means clustering to build the visual codebook. We call this approach [7] + SIFT. We assign to each pixel the code of the cluster in the visual codebook of its corresponding SIFT descriptor. We use a visual vocabulary size of 2000, and the parameter C of the regressor is fixed to $C = 1000$ using the validation set.

Second, we learn a Random Forest regression model for the vehicle densities following the model described by Fiaschi et al. in [8]. In this case, we integrate in our approach different features. We start using a simple feature: the normalized RGB values of the pixels ([8] + RGB Norm). The second feature we use is the Local Binary Pattern (LBP), using the VLFeat implementation [17] ([8] + LBP). The third feature type consists in a concatenation of the following filter responses ([8] + Filters): the gray-scale value of the pixel, the Laplacian of Gaussian filter response, the Gaussian gradient magnitude and the eigenvalues of the structure tensor of the image.

4.2 Vehicle Counting Evaluation

We start reporting the performance of the model HOG-1, which casts a MAE of 34.01. This confirms that: a) training in the PASCAL VOC is not adequate, due to the different nature of data distribution of the two datasets; b) the problem proposed by the TRANCOS dataset is very challenging.

Table 1 shows the results obtained for the rest of methods. We include both the MAE and GAME (for $L = \{1, 2, 3\}$) metrics. Let us start analyzing the MAE results. One first observes that the counting by detection HOG-2 drastically reduces the MAE of the HOG-1 to 13.29. This error can be considered equivalent to the one reported by the counting by regression model [7]+SIFT. With respect to the approaches following [8], we observe that the best results are obtained using normalized RGB features and the filter responses ([8] + RGB Norm + Filters). Another interesting conclusion is that the MAE of these state-of-the-art models is significantly higher than the previously reported performances of the same models addressing the crowd counting problem in other datasets. Again, this reveals that the problem proposed by the TRANCOS dataset is really challenging.

We can conclude that the best approach in terms of MAE is the HOG-2. However, the GAME metric shows that this is not true. We have observed that the HOG-2 approach casts multiple wrong detections, which contribute to improve the count of the objects, but they are actually false positives. The GAME metric is able to overcome this limitation, because it penalizes the wrong localizations of the object counts.

Observe Table 1 GAME columns for a clear comparison of the different methods. First, as it was expected, the higher L the higher the error of the models. The best performance is now systematically obtained by [7]+SIFT. Furthermore, in Figure 4 we can see that it is the HOG-2 the one suffering the biggest increment of the error. Also, for $L > 2$ all the counting by regression models improve the results of HOG-2. Our results show that [7]+SIFT is the best approach counting and localizing the vehicles in the images.

	MAE = GAME(0)	GAME(1)	GAME(2)	GAME(3)
[8] + RGB Norm	20.25	22.57	26.78	29.54
[8] + LBP	19.98	23.15	28.04	31.19
[8] + RGB Norm + LBP	20.15	22.66	27.04	29.97
[8] + Filters	17.77	20.14	23.65	25.99
[8] + RGB Norm + Filters	17.68	19.97	23.54	25.84
[7] + SIFT	13.76	**16.72**	**20.72**	**24.36**
HOG-2	**13.29**	18.05	23.65	28.41

Table 1. Vehicle Counting Results. We report the MAE and GAME metrics.

5 Conclusions

We conclude that the TRANCOS dataset introduces a very challenging and novel problem. The experimental evaluation proposed sets a new benchmark. With the novel GAME metric proposed, we overcome some of the limitations of the traditional MAE for object counting solutions, and provide a more precise evaluation where both the localization and the count are taken into account simultaneously.

Acknowledgements. This work is supported by projects SPIP2014-1468, CCG2013/EXP-047, CCG2014/EXP-054, TEC2013-45183-R and IPT-2012-0808-370000.

References

1. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC)challenge. IJCV **88**(2) (2010) 303–338

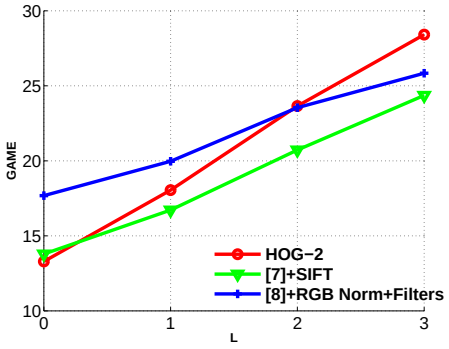


Fig. 4. GAME metric evolution with respect to L.

2. Guerrero-Gomez-Olmedo, R., Lopez-Sastre, R.J., Maldonado-Bascon, S., Fernandez-Caballero, A.: Vehicle tracking by simultaneous detection and viewpoint estimation. In: IWINAC. (2013)
3. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. IJRR (2013)
4. Caraffi, C., Vojir, T., Trefny, J., Sochman, J., Matas, J.: A system for real-time detection and tracking of vehicles from a single car-mounted camera. In: ITS Conference. (2012)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
6. Sudowe, P., Leibe, B.: Efficient use of geometric constraints for sliding-window object detection in video. In: ICVS. (2011)
7. Lempitsky, V., Zisserman, A.: Learning to count objects in images. In: NIPS. (2010)
8. Fiaschi, L., Köthe, U., Nair, R., Hamprecht, F.A.: Learning to count with regression forest and structured labels. In: ICPR. (2012)
9. Chan, A.B., Liang, Z.S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: Counting people without people models or tracking. In: CVPR. (2008)
10. Lu, W., Wang, S., Ding, X.: Vehicle detection and tracking in relatively crowded conditions. In: IEEE International Conference on Systems, Man, and Cybernetics. (2009)
11. Jun, G., Aggarwal, J.K., Gökmen, M.: Tracking and segmentation of highway vehicles in cluttered and crowded scenes. In: IEEE Workshops on Applications of Computer Vision. (2008)
12. Tamersoy, B., Aggarwal, J.K.: Robust vehicle detection for tracking in highway surveillance videos using unsupervised learning. In: AVSS. (2009)
13. Chen, K., Loy, C.C., Gong, S., Xiang, T.: Feature mining for localised crowd counting. In: BMVC. (2012)
14. Arteta, C., Lempitsky, V., Noble, J., Zisserman, A.: Learning to detect partially overlapping instances. In: CVPR. (2013)
15. Selinummi, J., Seppala, J., Yli-Harja, O., Puhakka, J.A.: Software for quantification of labeled bacteria from digital microscope images by automated image analysis. Biotechniques (2005)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2) (2004) 91–110
17. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/` (2008)