2016 15th IEEE International Conference on Machine Learning and Applications

Towards Deep Learning Invariant Pedestrian Detection by Data Enrichment

Cristina N. Vasconcelos, Aline Paes, Anselmo Montenegro
Department of Computer Science
Universidade Federal Fluminense (UFF), Niteroi, RJ, Brazil
{crisnv, alinepaes, anselmo}@ic.uff.br

***Abstract*—Deep learning models have recently achieved the state-of-the-art results on a well-known pedestrian detection dataset. However, such images were obtained from open scenarios with fixed imaging geometry parameters, which may produce a network not suitable for detecting a person in more general settings, such as the ones found in surveillance systems. As gathering and annotating data is a highly expensive manual task, we propose a methodology for artificially augmenting the positive training set with automatically generated local image affine and perspective transforms. Furthermore, to enrich the variability of background images, we include to the negative training set images that resemble human figures automatically obtained by the proposed methodology over images from commonly found surveillance scenarios. Extensive results show that by providing the enriched data as the input to a Convolutional Neural Network it is possible to precisely detect pedestrians in a number of public datasets. The data enrichment proposed here may also be used in other detectors based on supervised learning architectures, as the process is independent from the learning algorithm employed.**

I. Introduction

Pedestrian detection is a key Computer Vision problem in a number of high-level applications such as car safety and surveillance systems. It belongs to the large family of Pattern Recognition tasks, which in turn have witnessed significant improvements in the recent years, mainly due to the advent of deep learning techniques [1], particularly the Convolutional Neural Networks (CNN) [2].

Although the general neural network algorithms exist since the 1960s, only in the recent few years largely successful results have been obtained within these techniques, mostly due to the development of Graphics Processing Units (GPUs) and efficient training algorithms [3].

The broad interest in developing robust image analysis solutions motivated a number of groups to make their annotated datasets publicly available. This is the case of the pedestrians images datasets Caltech [4], Inria Person Dataset [5] and Kitti [6] and datasets encompassing an increasing number of different classes of objects like the ImageNet [7] dataset. With similar motivation, deep neural nets architectures such as AlexNet [2], and their corresponding learned weights are made available, which have contributed to avoid building datasets and networks settings from scratch.

The authors would like to thank NVidia for the donation of the GPUs used in the experiments and Brazilian Research Agencies CNPq and FAPERJ.

Recently, pedestrian detection through CNN has produced state of the art results in the Caltech public dataset [4]. The images there present open scenarios and were obtained by a camera placed on the top of a driving car.

Our hypothesis is that the geometry of such camera imposes a limitation on the kind of pedestrian silhouettes that can be used in the detection process. Therefore, one question that arises is how far training a deep architecture with such samples does generalize to cameras positioned in more general configurations.

Moreover, training a CNN is equivalent to estimating millions of parameters which in turn requires a very large number of *annotated* images [8]. Besides the annotation process being a laborious and time-expensive task, the training set may be a heavily biased subset of the real space. Thus, it is quite important to have mechanisms that not only automatically increase the training set with new samples but also augment the data with new features that add variability to the class of target objects of the task at hand, in our case the pedestrians.

This paper focuses on increasing pedestrian detection accuracy on surveillance datasets by enriching training samples through fully automatic procedures. Specifically, we present a methodology to fine-tune the state of the art solution, augmenting already labeled databases in two directions (as detailed in Section II): (1) artificially simulating imaging geometry variations; (2) enriching the knowledge about commonly found surveillance scenarios.

We investigate how the CNN trained with samples taken from the fixed settings [4] does generalize to pedestrian detection in surveillance datasets and show how our proposal improves the CNN invariance to imaging geometry and background changes (as detailed in Section III).

A. Related Work

To the best of our knowledge, the most recent work tackling a straightforward CNN in pedestrian detection task is [9]. That work explored the parameters design space of the AlexNet [2] deep network, achieving the best known performance on Caltech pedestrian dataset [4]. Consequently, in this work we build our method on top of the best architecture and methodology reported thereby. However, a question that arises is whether standard pedestrian datasets are rich enough to allow the detection of intrinsic camera variations presented

978-1-5090-6167-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICMLA.2016.169

in surveillance videos. To address this question, we take a step further, as we propose to augment the training set of the Caltech dataset devised in [2] with pedestrian and background variations, and see how the networks trained on both "initial" and enriched datasets behave on other pedestrian datasets.

Previous works focusing on other tasks also introduced additional examples representing variations of images to standard datasets, in order to increase the performance of machine learning algorithms. For example, this has become a standard practice when detecting handwritten digits in the well-known MNIST dataset, where it is usual to apply affine displacement field to images to obtain their transformed versions encompassing translations and rotations [10].

Including artificial training samples is also useful to take into account a prior knowledge about a specific task. In [11] this technique is thoroughly discussed in a number of scenarios, including using prior knowledge about invariances to generate "virtual" examples when learning SVMs.

In this work we also contribute with training set enrichment methodologies but by adding background and affine and perspective variations to a standard pedestrian detection dataset. As far as we know, this is the first work following the proposed image transforms to enrich a CNN detector.

II. Method

This section describes our method designed to augment both the CNN invariance to image formation geometry and its knowledge about the surveillance context.

A. Imaging Geometry

It is well known that the image formation process is influenced by intrinsic and extrinsic parameters of the camera used. The images in the Caltech dataset [4], from which all the training samples in [9] were obtained, were captured according to fixed automotive settings. (640×480 pixels, fov = $27°$, focal length at $7:5$ mm and a projection plane parallel to the human figure).

One cannot assume that these exact values correspond to general surveillance settings, neither by the defined projection geometry, nor by the intrinsic parameters described. We question that a CNN trained over such image samples may not generalize well when faced with images taken in differing conditions.

Ideally, new photos taken from varying camera extrinsic and intrinsic parameters would enforce the knowledge of the pedestrian detector towards an invariance to imaging geometry, but it would require huge manual effort.

Artificially producing the variations requires a global model of the scene, but finding it amounts to inferring the entire 3D structure of the scene.

With that in mind, even though it is not possible to correctly simulate the full range of existent variations through image domain transformations, we investigate how a local model can contribute to enlarge the training set by artificially introducing small perturbations into each original training sample through affine and perspective warps.

The affine transform is a good approximation for small planar patches to the image plane undergoing an arbitrary translation and rotation about the optical axis, and small rotations about an axis orthogonal to it. While in affine transformation, all parallel lines in the original image will still be parallel in the output image, this is not guaranteed by the perspective transforms that only guarantees to keep the straight lines after the transformation. Thus, our projective model simulates small deformations along the scene depth.

To produce a random affine transform and keep it local, we map three anchor points from the border of the pedestrians bounding boxes and disturb them by random vectors of at most 2 pixels. Three non-collinear 2D points represented in homogeneous coordinates are necessary as the affine transform is well defined by a 2×3 matrix A, which has 6 degrees of freedom. Thus, a transformed point $p\prime$ is computed as $p\prime = Ap$ where p is the corresponding point in the original image.

We use a subclass of perspective transforms called homographies which are 2D perpective transforms defined on points restricted to a plane. We create the random homography similarly to the affine case, but instead of three anchor points we require four anchors to define the 8 degrees of freedom of a homography. Our points are represented in homogeneous coordinates and the transform is defined by a 3×3 matrix H which is defined up to a scale. Similarly to the affine case, a point $p\prime$ in the transformed image is computed as $p\prime = Hp$ where p is the corresponding point in the original image.

Both positive and negative artificial samples, for affine and perspective are computed perturbing the same anchor points.

The local transformation may alter the position of the tip of the head and the foot base pixels. But, for positive samples, keeping fixed those two reference points is very important in order to do not loose the registration made during original annotation of the pedestrians bounding boxes that ultimately induces the good localization capacity of the trained network. The requirement of a registered dataset is observed in training CNNs for object detectors, but not for the tasks that do not demand precise localization within the image. Thus, for both affine and perspective transforms, after the random transformation is applied to the image, the referred two points are mapped again to the same positions within the artificial bounding box as they occupy in the original training sample.

B. Surveillance Common Backgrounds

The backgrounds observed in Caltech dataset [9] mostly present sky, streets, vehicles, outdoors, building facades and nature elements. Those are not the typical objects retrieved by surveillance cameras from shops, banks, airports, and others. We propose to enrich the knowledge of network by augmenting the background training samples for this broad context.

With this goal, we selected sixteen synsets of the ImageNet [7] database (detailed in Table I).

Each image represents a possible background or an object commonly found in surveillance scenes, and it is sampled ten times in random positions and scales while keeping fixed

TABLE I: Selection of Imagenet Synsets for Surveillance: the 19954 images were randomly cropped, but maintaining the training bounding box aspect ratio. These are further evaluated to produce $2 * 905$ background hard-cases (flipped copies are also created)

Synsets	#images	#hard cases	Synsets	#images	#hard cases	Synsets	#images	#hard cases	Synsets	#images	#hard cases
Furniture	2138	169	Door	1190	9	Wall	516	5	Case	1243	12
Floor	1207	22	Fabric	1698	30	Structure	1190	9	Carpet	944	10
Paving	1235	9	Field	1289	14	Shopping	1260	48	Shop	1236	83
Station	973	8	Concrete	1315	9	Artifact	1249	447	Plant	1271	21

the pedestrian bounding box aspect ratio. A few images with resolution smaller than the pedestrian standard bounding box (64×128 pixels) were discarded. The set of new background samples created contains around two hundred thousands images.

The set of generated samples is further evaluated to create a hard cases subset. For that, each sample is classified using the CNN from [9]. Next, we discard all the samples classified as background assuming that they can be considered easy to classify by the already trained model. A final verification is made with the remaining candidates, as the Imagenet Synsets may contain a pedestrian that can be cropped correctly during the random process by coincidence and should be discarded. Figure 1 illustrates some hard-cases produced.

C. Convolutional Neural Network Architecture

We used the CNN architecture known as Alexnet [2] with the modifications proposed in [9]. Figure 2 exhibits the AlexNet architecture as used in this paper.

It comprises eight learned layers, where the first five are convolutional layers, some of them yielding max-pooled and normalized outputs, the sixth and seventh layers are fully-connected and the last one is a Softmax output. As in [9], we modified the last layer to produce a binary classifier and fine-tuned it from the weights learned with the ImageNet dataset [7], to avoid overfitting. All the parameters are kept as in [9], except for the batch size, which here is of 256 examples.

III. Experimental Results

Tests were made on public datasets in order to evaluate how does the Alexnet architecture fine-tuned over the Caltech dataset generalize to surveillance and how our proposal improves such generalization.

We took advantage of the visual NVIDIA Deep Learning GPU Training System (DIGITS) [1] software to perform the experiments, which in turn uses the Caffe Deep Learning framework [12] to in fact define, As usual, to enforce the generalization criteria, neither the surveillance datasets nor the Caltech's test samples were used during training, or to execute the data augmentation procedures proposed in this paper. The single training parameter altered from [9] was a batch size of 256 examples, but all others kept fixed. The models presented were chosen as those presenting the highest accuracy on Caltech's validation set within 15 training epochs. No other architecture, or set of parameters was evaluated as it was not our main concern, that is, how the augmented data can improve pedestrian detection given a CNN. The experiments were executed with a GeForce GTX Titan X (12GB RAM).

First, we fine-tuned a CNN with the Caltech's original training samples and applied it over the Caltech's test subset and a set of surveillance datasets in order to produce baseline results for the subsequent experiments comparison observing false positive rate (FPR) and miss rate (MR). Next, we proceed with the experiments conducted from our two proposals, i.e., augmenting (1) the pedestrian data with geometry invariance and (2) the background with common surveillance images.

With respect to the simulation of image geometry variations, a training set was made augmenting each original image with three artificial samples by random affine transforms (detailed in subsection II-A); a second set by three random perspective transforms; and a last one including both affine and perspective samples (results in Table II).

By introducing new samples with the affine transformation, we were able to improve the miss rate of three datasets, compared to the original network, trained only with the unchanged Caltech examples. Three other datasets had their miss rate improved with the additional perspective transformation samples. Five other datasets presented an improvement on the miss rate when both affine and perspective transformations were used to generate additional pedestrian samples. Thus, we were able to improve the miss rate of eleven of the thirteen datasets tested in this paper, by using one or both the transformations. Furthermore, eight of the thirteen datasets had their false positive rate also decreased and nine of them also presented an improvement on the F1 measure, showing that with additional pedestrian examples we can increase the trade-off between predictive precision and recall in the pedestrian detection problem.

The datasets that still continued with overall better results in the original approach (trained without the new samples) are those filmed with too elevated/low angles, which are arguably not well represented with the new samples generated from affine and perspective transforms. This may indicate that there is still room for more improvement by adding more radical transformations, yet one should be careful to not introduce unreal distortions into the dataset.

Regarding the background generalization (Table III), an initial training was made containing all the original samples taken from Caltech plus approximately two hundred thousand new negative samples produced by sampling our selection of Synsets from Imagenet (Section II-B). Including all the

[1] https://developer.nvidia.com/digits

Fig. 1: Hard cases automatically created as new negative training samples

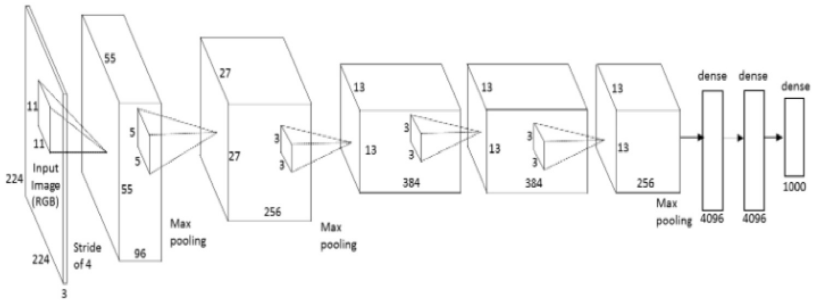

Fig. 2: AlexNet Architecture [2]. In the present work, the last layer has size 2, instead of 1000.

TABLE II: False Positive Rate (FPR), Miss Rate (MR) and F-Measure (F1) achieved when augmenting Caltech's training samples towards image geometry invariance (values are scaled by 10^4). The bold values are the best ones for each dataset.

Dataset name	People	Original [9]			+affine			+perspective			+affine+perspective		
		FPR	MR	F1	FPR	MR	F1	FPR	MR	F1	FPR	MR	F1
Caltech test set[4]	16264	**83**	106	9903	91	101	9902	101	70	9912	107	**69**	**9909**
TownCentre[13]	7148	**185**	305	**9405**	220	344	9307	205	**300**	9362	223	314	9315
Venice-2 [14]	7141	69	327	9697	**65**	387	9675	74	364	9669	75	**320**	**9690**
ADL-Rundle8 [15]	6783	88	771	9404	**86**	748	9422	89	**699**	**9442**	94	702	9428
ETH-PedCross2 [16]	6263	**106**	1975	8814	122	1970	8803	116	1965	8812	108	**1949**	**8828**
ETH-Bahnhof [16]	5415	138	89	9656	125	149	9654	**124**	93	**9683**	137	**86**	9661
ADL-Rundle6 [15]	5009	114	**564**	**9510**	**99**	696	9465	105	675	9466	115	671	9452
PETS09-S2L1 [17]	4476	**42**	**132**	**9831**	51	205	9771	47	168	9800	50	189	9783
ETH-Sunnyday [16]	1858	114	169	9677	126	**112**	9681	110	146	9696	**108**	135	**9707**
TUD-Stadtmitte[18]	1156	676	**0**	8765	**574**	10	**8927**	649	**0**	8809	581	**0**	8920
KITTI-13 [6]	762	527	76	8843	**495**	91	**8895**	562	**60**	8785	574	**60**	8762
KITTI-17 [6]	683	**186**	96	**9572**	209	**24**	9564	203	72	9551	254	**24**	9477
TUD-Campus [19]	359	503	299	9012	**380**	**149**	**9279**	390	261	9206	421	224	9177

new background samples creates a distortion in the positive-negative samples ratio. Such imbalance induces an improvement in the false positive rate, but sometimes at a high cost of also reducing the true positive rate. In a second experiment, we added to the set of background examples about two thousand hard cases, selected from Imagenet collection. We used the four networks trained with Caltech, with and without the automatically generated pedestrian samples, whose results are in the first line of Table II. With that, we would like to insert in the set of examples a trade-off between background and pedestrian detection.

In this case, we were able to improve the F1 measure of twelve of the thirteen test sets, with at least one of the networks trained with additional negative and positive samples. Only the $PETS09$ [17] dataset, which shows eight walking pedestrians in rather unusual patterns, still had better results with the network trained with only the original Caltech samples.

IV. Conclusions

Inspired by the requirement of huge amounts of labeled data for training a Deep CNN and by previous successful attempts of increasing invariance properties of supervised learning algorithms through annotated data augmentation [11], we presented a methodology for automatically generating new samples for training a pedestrian detector based on two techniques. We artificially warped the Caltech's samples with controlled affine and perspective transformations to simulate image geometry variations. At the same time, the pedestrians head and foot positions inside the bounding boxes are maintained fixed protecting the ground truth registration.

TABLE III: False Positive Rate (FPR) and Miss Rate (MR) when augmenting background data from [4] with (ALL) cropped samples from Imagenet synsets proposed selection or only hard cases (HC) for the background data augmentation proposed (values are scaled by 10^4). The bold values are the best ones for each dataset.

Dataset name		Original positive samples			+affine			+perspective			+affine+perspective		
		FPR	MR	F1	FPR	MR	F1	FPR	MR	F1	FPR	MR	F1
Caltech test set[4]	ALL	71	160	9882	90	**91**	**9907**	72	124	9899	**66**	142	9894
	HC	**75**	136	9892	86	93	9908	84	88	9912	87	**84**	**9912**
TownCentre[13]	ALL	**11**	4250	7275	15	**4107**	**7380**	13	4312	7222	15	4193	7312
	HC	**34**	2025	8790	48	**1848**	**8865**	41	1954	8817	41	2148	8698
Venice-2 [14]	ALL	19	1423	9196	20	**993**	**9439**	**15**	1198	9333	18	1061	9405
	HC	**27**	776	9542	75	5250	6312	28	**759**	9549	41	593	**9614**
ADL-Rundle8 [15]	ALL	**9**	3058	8177	18	2568	8488	16	2669	8426	16	**2549**	**8504**
	HC	**29**	2052	8794	413	5564	5443	33	1964	8838	35	**1856**	**8899**
ETH-PedCross2 [16]	ALL	**7**	5379	6315	16	**4549**	**7043**	13	4652	6959	15	4749	6874
	HC	31	4004	7472	548	5842	5502	44	**3775**	7637	**29**	3782	**7643**
ETH-Bahnhof [16]	ALL	**37**	860	9469	50	**535**	9615	39	624	9592	40	564	**9621**
	HC	89	**275**	9666	718	1533	7813	**67**	340	9679	80	279	**9683**
ADL-Rundle6 [15]	ALL	**5**	3449	7907	10	3081	8162	6	3338	7987	11	**2960**	**8243**
	HC	**25**	**1339**	**9238**	276	3176	7662	33	1460	9154	30	1480	9148
PETS09-S2L1 [17]	ALL	**5**	**749**	**9599**	6	769	9584	6	833	9551	6	824	9555
	HC	**14**	**339**	**9792**	3670	3456	3765	19	414	9741	17	462	9722
ETH-Sunnyday [16]	ALL	17	3135	8106	18	**1955**	**8878**	22	2264	8677	**16**	2079	8807
	HC	41	1146	9306	344	4736	6291	45	865	6291	**37**	**719**	**9549**
TUD-Stadtmitte [18]	ALL	180	804	9220	173	**678**	**9302**	**133**	1190	9100	178	752	9253
	HC	431	**84**	9135	**273**	5929	5353	293	177	9335	296	125	**9357**
KITTI-13 [6]	ALL	180	861	9139	199	6665	9203	**167**	634	9288	177	**574**	**9230**
	HC	322	408	9078	**240**	2311	8164	278	363	**9193**	303	**317**	9164
KITTI-17 [6]	ALL	56	1435	9109	39	1077	9348	28	1077	9372	**23**	**1053**	**9397**
	HC	85	789	9413	333	1866	8323	**51**	526	**9623**	68	**478**	9614
TUD-Campus [19]	ALL	318	1604	8588	246	**1306**	**8876**	**236**	1642	8699	246	1716	8638
	HC	431	**896**	8809	**195**	4888	6462	339	1082	8852	329	933	**8950**

In another direction, to enrich the knowledge of the network about surveillance common scenarios, we revisit surveillance vocabulary to produce a big set of new background images that are further evaluated to produce a selection of hard cases for fine-tuning the pedestrian detector.

Both miss and false positive ratios were improved in a number of experiments but identifying a balance between them is not a trivial task. The ratio of negative over positive samples used in the training was strongly influenced towards reducing FPR but augmenting MR in cases where more negative samples were added. A natural future experiment is to include in our methodology a procedure to discard negative samples from the original training set in order to restore the samples balance. On the other hand, the miss rate in the experiments including perspective artificial samples but keeping negative-positive balance, achieved significant results that validate the proposed method in most of the tested surveillance datasets.

Finally, the data enrichment proposed here may also be adopted for preparing new training samples for other object detectors based on supervised learning.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, p. 436444, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS 25*, 2012, pp. 1097–1105.

[3] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[4] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," 2009.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," pp. 886–893, 2005.

[6] A. Geiger, "Are we ready for autonomous driving? the kitti vision benchmark suite," pp. 3354–3361, 2012.

[7] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," pp. 248–255, 2009.

[8] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, "Deep image: Scaling up image recognition," *CoRR*, vol. abs/1501.02876, 2015.

[9] J. H. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," pp. 4073–4082, 2015.

[10] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," pp. 958–962, 2003.

[11] D. Decoste and B. Schölkopf, "Training invariant support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 161–190.

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[13] B. Benfold and I. Reid, "Guiding visual surveillance by tracking human attention," 2009.

[14] ——, "Guiding visual surveillance by tracking human attention," 2009.

[15] L. Leal-Taixé, A. Milan, I. D. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *CoRR*, vol. abs/1504.01942, 2015.

[16] A. Ess, B. Leibe, and L. J. V. Gool, "Depth and appearance for mobile scene analysis," pp. 1–8, 2007.

[17] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," 2009.

[18] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," 2010.

[19] ——, "People-tracking-by-detection and people-detection-by-tracking," 2008. [Online]. Available: http://www.mis.tu-darmstadt.de/node/382