SYNTHETIC TRAINING IN OBJECT DETECTION

*Osama Khalil, Mohammed E. Fathy, Dina Khalil El Kholy, Motaz El Saban**,
Pushmeet Kohli, Jamie Shotton and Yasmine Badr

Microsoft Advanced Technology Labs, Cairo, Egypt

ABSTRACT

We introduce new approaches for augmenting annotated training datasets used for object detection tasks that serve achieving two goals: reduce the effort needed for collecting and manually annotating huge datasets and introduce novel variations to the initial dataset that help the learning algorithms. The methods presented in this work aim at relocating objects using their segmentation masks to new backgrounds. These variations comprise changes in properties of objects such as spatial location in the image, surrounding context and scale. We propose a model selection approach to arbitrate between the constructed model on a per class basis. Experimental results show gains that can be harvested using the proposed approach.

***Index Terms*—** Synthetic training, Object detection

1. INTRODUCTION

Most object detection and localization methods require the availability of a set of training images and ground-truth annotations for an object class. The contrcuted object model accuracy strongly depends on the amount of training data and the variation within it. Unfortunately, gathering and annotating training images is an expensive and time consuming task. In this paper, we investigate different approaches for synthetically augmenting the available training set. While synthetic augmentation methods have been studied before, they mostly relied on global modifications to the real training set, e.g. scaling, rotation,... On the other hand, the proposed methods in this paper target a complete change in object background to boost variability in the training set. These methods require availability of the segmentation mask in the images that constitute the *original* training set. We use the PASCAL Visual Object Category Challenge (VOC) 2009 [1] dataset for our experiments which provides segmentation masks for some of the training images. Using its segmentation mask, we can relocate the object at different positions in different images as illustrated in Figure 1. The main contributions of this paper are: a) Proposing a number of different synthetic image generation methods that go beyond simple global scaling, b) Using model selection techniques to arbitrate between models generated by different training datasets constructed from the proposed synthetic methods on a per class basis and c) Establishing that synthetically generated images can partially replace manually labeled ones for object detection training.

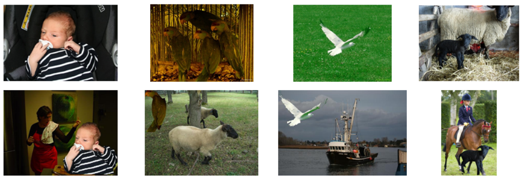

Fig. 1. Example synthetic image generation. Top row depicts original images and bottom row shows synthetic images.

2. RELATED WORK

There have been several recent attempts to make use of synthetic training examples to enhance test performance for many learning-based approaches [2, 3, 4, 5, 6, 7]. According to Nonnemaker [8], synthetic training generation can be performed in the original sample space [2, 9, 3, 4], parameter space [8] or feature space [10]. For an example of sample space, [3, 11] utilized global geometric and photometric transformations to create synthetic images. In [2], the synthetic training examples were obtained by placing an artificial 3D person mesh on top of random real images. Marin et al. [12] investigated the problem of using a pedestrian appearance model learnt from virtual environments for detection in real scenes. The idea of synthesizing images from an original dataset has also been suggested in different contexts, such as constructing a more challenging object recognition dataset to better mimic real world issues as in the work by Kinnunen et al on the Caltech-101 dataset [13]. More recently, Nga et al. [14] exploited synthetically generated videos.
[9] used synthetic training face images to train a face detector through learning a 3D model for each person and using it to synthesize images by changing pose and illumination conditions. Similarly, the idea of synthetic training data usage has been experimented with for cursive handwritten recognition [4] by using a perturbation model in the generation pro-

[*] Author performed the work while on leave from Cairo University, Faculty of Computers and Information.

cess. LeCun et al. [15] added different sources of variations such as object perturbation, superposition on a complex background and adding distractor objects. While this is similar in part to our work, the authors augmented the dataset without investigating how this could affect performance on trained models. In addition, their approach is limited to training sets in which the objects are imaged under constrained settings (mainly a uniform background). Alternatively, [8] uses a parameter space approach to generate new training samples for isolated character recognition. Finally, Jiang et al. [10] generated synthetic training data by sampling the feature space distribution for a signer-independent sign language recognition system. Among the different alternatives for synthetic data generation, the one that operates in the original signal (image) space is the most appealing to investigate as it has a clear semantic meaning, as opposed to manipulation in the feature space. Besides, using a parametric space generation approach is not practical as there are no such parametric models capable of generating natural images as opposed to the more restricted case of characters [8].

In this paper, we investigate methods completely replacing the object background while minimizing artifacts. There has been previous work in the area of attempting to produce realistic looking images when inserting novel objects on an image such as [16], [17], [18] and [19]. The main drawback of these approaches is that they still rely on heavy user interaction and the knowledge about the proper scale of the inserted object.

3. RELOCATION METHODS

The goal of having multiple relocation strategies is to accommodate for every object class properties. Each relocation strategy aims at emphasizing a certain type or types of backgrounds not well represented in the initial dataset.

3.1. Relocation within Same Image

Given a particular training image, we check sequentially all objects inside it and try to relocate them (one-by-one) to a different location within the same image. As previously mentioned, we assume the availability of an object level segmentation mask such as the one supplied with the VOC segmentation challenge. We place objects on an area having a "large" background to foreground ratio. Additionally, since we cannot guarantee to find an area large enough to host the displaced object, we may scale down the object following a specific scaling schedule. We use a sliding window over the target image to identify a box to place the object. We experiment with different sizes of the target bounding box, all made relative to the original object bounding box. Among all checked boxes with their possible scaling, we select the one maximizing the background to foreground ratio.

3.2. Relocation to Images with Co-occurring objects

In this method, we relocate an object O to a background area only in images with objects highly co-occurring with O. We identify the top co-occurring object class and denote that class as C_Z. We rank all images belonging to C_Z based on similarity of the object's scale with respect to its class, denoted by S_{Rel_O}, to that of the highest co-occurring object in the target image denoted by S_{Rel_K}. The relative scale of a certain object O to its class is defined as follows:

$$S_{Rel_O} = \frac{S_O - \overline{S}_{C(O)}}{max(S_{C(O)}) - min(S_{C(O)})} \quad (1)$$

where S_O is measured by the area of the bounding box of object O, $\overline{S}_{C(O)}$ is the average bounding box area of O's class, $max(S_{C(O)})$ is the maximum and $min(S_{C(O)})$ is the minimum area of all object instances in O's class. Similarity is computed using the monotonically decreasing function in Equation 2 and we insert the source object in the most similar image.

$$Sim = \frac{1}{1 + (S_{Rel_O} - S_{Rel_K})^2} \quad (2)$$

3.3. Relocation to Images Different in Appearance

The goal of this method is to relocate the object to a background sufficiently different in appearance from the source background. We use normalized histograms of oriented gradients to describe the appearance of the background. For every image, we compute a feature vector concatenating 3 histograms of 9 gradient bins each. This feature vector describes gradient densities in the background of each of the 3 geometric components of the image; namely sky, ground and vertical components (detection of geometric structure is described later). We, then, cluster the feature vectors into *n* clusters (we used *n* = 10, 20 and 30). Now, given a source object, a target cluster for relocation is selected as follows: we sort the clusters' centroids in order of their L1 distance from the centroid of the cluster of the source image. We select the median cluster centroid as the target cluster for relocation. Once a class of backgrounds has been decided upon, images within that class are penalized based upon their divergence from the source image in two respects: the geometric structure and the illumination map.

Geometric Structure. Using geometric context proposed in [19], we classify every pixel as sky, ground or vertical plane pixel. A 3-bin histogram of the percentage of each geometric component in the image is computed.

Illumination Map. We follow the method suggested in [19] to relocate objects to backgrounds most matching their illumination. For each image, we compute a 3-bin illumination histogram reflecting the illumination density in CIE Lab color space in each of the 3 geometric components (sky, ground and vertical). The calculated vector is normalized to have unit

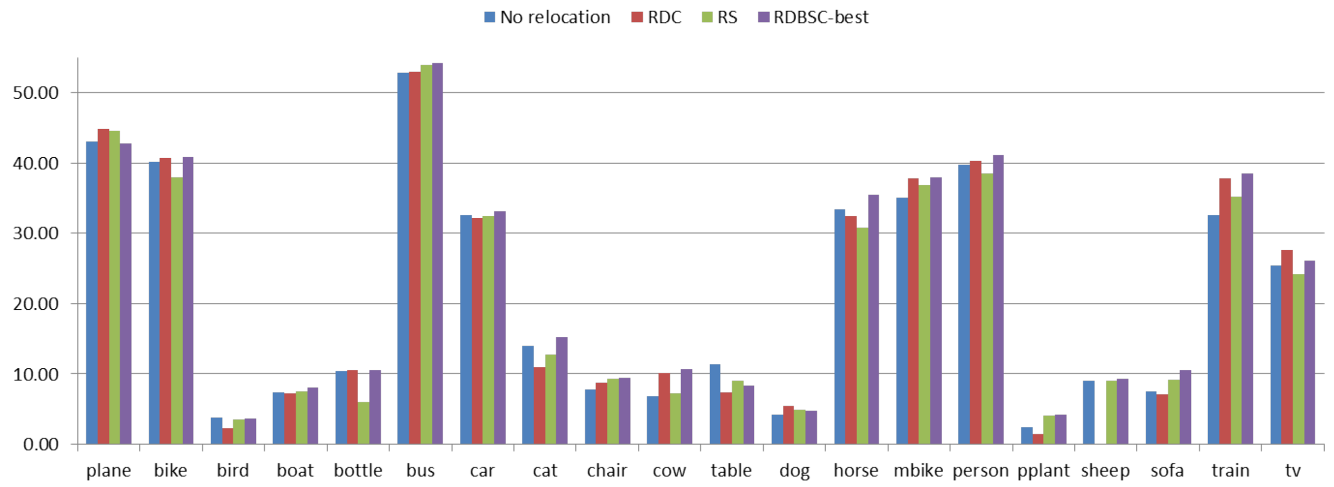

Fig. 2. Variation in performance of relocation methods across classes. RS, RDC, RDBSC refer to methods in sections: 3.1, 3.2 and 3.3 (with # of clusters chosen per class to be the highest achieving among n = 10, 20 and 30) respectively.

magnitude. Each image, in the selected cluster, gets scored by the sum of the weighted distances of illumination map vector and geometric structure vector from the source image. An image minimizing this score gets selected for relocation. Within the selected image, we search for the rectangular area with the most background and resize the source object if necessary.

4. SELECTION OF RELOCATION METHOD

We propose that selection among relocation strategies (including no relocation) should be performed, per class, of the method achieving the highest average precision on average in 3-fold validation dividing the training set into 70% for training and 30% for validation. Figure 2 shows results of our model selection experiment using deformable parts based model [20] as baseline detector. Variations in performance could be attributed to the interplay between the variation in the role played by the background in the detection task for every class and how each relocation method responds to such class specific properties. To give an idea of such claim, we studied (omitted from paper for space reasons) the comparative behavior of relocation to images with different appearance (denoted RDBSC) in relation to the baseline (no relocation) using VOC 2009 train set for training and val set for validation. We notice that stability of the background in a certain object class suggests that the background is clue for detection of that class and thus contraindicates relocation to backgrounds significantly different from those in the dataset. The bird class is a good example of such phenomenon where the background is mostly a blue sky and thus relocation can be a source of noise introducing unlikely variations to the dataset and hence the degradation of performance by RDBSC from the baseline. On the contrary, classes exhibiting high variability in their backgrounds show better performance using relocation to sufficiently different backgrounds. Sofa class would be our example in such situation. The rationale here is that such relocation technique provides more samples from the space of backgrounds for the learning algorithm. Finally, we note that the low proportion occupied by the background in the bounding box on average reduces the space of available background variations and in turn a sufficiently large training set (like VOC 2009) would very likely cover the needed background variations. In such case, relocation would be of little benefit to the detector and could even be a source of bias in the training set towards the foregrounds of the relocated examples since the synthesized examples introduce very little variations. TV monitor class would be a good example in such case where most object instances are box shaped and thus occupying most of the bounding box and gain achieved with RDBSC is 0%.

5. EXPERIMENTS

Experiments are designed to verify two hypotheses:
Hypothesis 1. *Synthetic training can boost object-detectors' performance through complementing the available training set with novel variations.*
For this purpose, we produced 2 sets of results using deformable parts-based model [20] as baseline detector. Our results are obtained on twenty object classes using PASCAL VOC 2009 [1] dataset. All results were obtained by partitioning trainval set (7054 images) into 50% for training and 50% for validation. The results presented were cross validated twice. Relocation methods used were selected per class as

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
BL	41.23	46.4	5.51	3.03	20.68	45.24	36.07	14.26	11.62	11.86	5.77	5.23	36.75	32.73	43.23	3.36	15.48	12.66	35.14	25.53
Synth	41.21	47.99	5.51	8.01	19.54	46.67	34.9	16.75	11.94	13.59	5.77	6.41	37.96	33.16	42.68	4.23	9.32	14.04	36.62	23.34

Table 1. Performance comparison of methods (referred to as Synth) selected as in section 4 on VOC 2009 dataset and deformable parts-based model as baseline detector (referred to as BL).

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	Average
BL-%set	41.89	44.495	4.67	4.145	20.175	46.255	33.38	13.92	11	8.1	8.08	5.025	36.2	30.28	42.91	3.55	10.95	13.57	35.865	23.975	21.92125
Synth-%set	39.17	47.44	4.67	3.03	20.285	47.305	32.83	13.59	11.7	10.15	8.08	4.755	36.935	31.14	42.36	5.69	14.9	14.45	34.96	24.41	22.4
BL-full set	41.23	46.4	5.51	3.03	20.68	45.24	36.07	14.26	11.62	11.86	5.77	5.23	36.75	32.73	43.23	3.36	15.48	12.66	35.14	25.53	22.6

Table 2. Hypothesis 2 experiment. BL-full set, BL-%set and Synth-%set refer to models trained on full training set, on a proportion of the set and on a synthetically augmented proportion of the set respectively.

explained in section 4. Table 1 shows our selected models outperform the baseline detector in 8 classes with over 1.5 AP points and in 3 more classes with less than 1 AP point.

Hypothesis 2. *Synthetic training can partially replace manual collection and annotation of huge datasets.* For this purpose, we produced three sets of results; baseline detector trained on the full training set, on a proportion of the training set and on the same proportion of the training set augmented using our synthetic approaches. Full training set means 50% of VOC 2009 trainval set. The proportion selected for the latter two results sets is determined so that the total number of training examples after synthetic augmentation is equal to the number of training examples in the fully manually annotated training set. The images removed were randomly obtained and results were cross validated in 2 folds. Table 2 shows that our approach got accuracy comparable with using the full manually labeled set reducing the 0.7 AP points drop on average to only 0.2 AP points. We also managed to outperform the baseline detector trained with fully manually labeled set in 7 classes (out of 20) with gain greater than 1.8 AP points in 4 of them.

It is worth noting that the experiment showed peculiar results on 7 classes where using a proportion of the training set produced models that outperformed their counterparts produced through training on the full training set in spite of cross validating the results. Such peculiarity is probably attributed to the presence of few variations in the object class relative to the size of the training set and thus the larger set could have introduced bias. Finally, we note that model selection experiments as well as experiments on the full validation set have revealed relocation within the same image has little benefit for the majority of classes. The reason could be attributed to the lack of novel variations introduced by this technique.

6. CONCLUSIONS AND FUTURE WORK

We have presented several new methods for synthetically increasing the training data for object recognition in images. The aim is to expose the model while training to variations aiming at better generalization on unseen data. We have wrapped the proposed synthetic data generation methods within a model selection framework since performance of the different synthetic methods varies across object classes. The conducted experiments have indeed shown that utilizing synthetic training data can boost recognition performance on many classes.

As for future work we want to investigate the effect of the proposed synthesization methods on other benchmark data and on possibly other tasks such as semantic segmentation. We explored some of the technical difficulties that might arise with other datasets such as the automatic generation of object segmentations in case these datasets provide only object bounding boxes or provide segmentation masks for a subset of the dataset. We conducted a pilot experiment comparing the performance of synthetic approach of relocation within the same image using only manual segmentation masks (436 images) to its counterpart using manual in addition to automatically generated segmentation masks (around 3000 images) using GrabCut [22] as a binary segmentation method. We observed performance variations (gains and degradation) using GrabCut masks across classes, the analysis of which is left for future work.

7. REFERENCES

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” *IJCV*, 2010.

[2] J. Yu, D. Farin, C. Krueger, and B. Schiele, “Improving person detection using synthetic training data,” in *ICIP*, 2010.

[3] J. Shotton, M. Johnson, and R. Cipolla, “Semantic texton forests for image categorization and segmentation,” in *CVPR*, 2008.

[4] T. Varga and H. Bunke, “Comparing natural and synthetic training data for off-line cursive handwriting recognition,” in *IWFHR*, 2004.

[5] J. Cano, J. Perez-Cortes, J. Arlandis, and R. Llobet, “Training set expansion in handwritten character recognition,” in *Proc. 9th SPR*, 2002.

[6] J. Sun, Y. Hotta, Y. Katsuyama, and S. Naoi, “Low resolution character recognition by dual eigenspace and synthetic degraded patterns,” in *ACM HDP*, 2004.

[7] D. Decoste and B. Scholkopf, “Training invariant support vector machines,” *ML*, 2002.

[8] J. Nonnemaker and H. S. Baird, “Using synthetic data safely in classification,” in *DRR*, 2009.

[9] B. Weyrauch, B. Heisele, J. Huang, and V. Blanz, “Component-based face recognition with 3d morphable models,” in *CVPRW*, 2004.

[10] F. Jiang, W. Gao, H. Yao, D. Zhao, and X. Chen, “Synthetic data generation technique in signer-independent sign language recognition,” *Pattern Recognition Letters*, 2009.

[11] V. Lepetit, P. Lagger, and P. Fua, “Randomized trees for realtime keypoint recognition,” in *Proc. CVPR*, 2005.

[12] J. Marin, D. Vazquez, D. Geronimo, and A. Lopez, “Learning appearance in virtual scenarios for pedestrian detection,” in *CVPR*, 2010.

[13] T. Kinnunen, J. Kamarainen, L. Lensu, J. Lankinen, and H. Klviinen, “Making visual object categorization more challenging: Randomized caltech-101 data set.,” in *ICPR*, 2010.

[14] Hang Nga and Keiji Yanai, “Improving person detection using synthetic training data,” in *ICCV*, 2011.

[15] Y. LeCun, F.J. Huang, and L. Bottou, “Learning methods for generic object recognition with invariance to pose and lighting,” in *CVPR*, 2004.

[16] J. Jia, J. Sun, C. Tang, and H. Shum, “Drag and drop pasting,” in *SIGGRAPH*, 2006.

[17] M. Johnson, G. Brostow, J. Shotton, O. Arandjelovic, V. Kwatra, and R. Cipolla, “Semantic photo synthesis,” in *Eurographics*, 2006.

[18] D. Hoiem, A. Efros, and M. Hebert, “Putting objects in perspective,” in *CVPR*, 2006.

[19] J. Lalonde, D. Hoeim, A. Efros, C. Rother, J. Winn, and A. Criminisi, “Photo clip art,” in *SIGGRAPH*, 2007.

[20] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *TPAMI*, vol. 32, no. 9, 2010.

[21] B. Russell and A. Torralba, “LabelMe: a database and web-based tool for image annotation,” *IJCV*, 2008.

[22] C. Rother, V. Kolmogorov, and A. Blake, “Grabcut: Interactive foreground extraction using iterated graph cuts,” *SIGGRAPH*, 2004.