# Synthetic Data Generation for Deep Learning in Counting Pedestrians

Hadi Keivan Ekbatani, Oriol Pujol and Santi Segui
*Faculty of Mathematics and Computer Science, University of Barcelona,*
*Gran Via de les Corts Catalanes, 585 08007 Barcelona, Spain*
*hadi.keivan.ekbatani@stu.fib.upc.edu, oriol_pujol@ub.edu, santi.segui@ub.edu*

Keywords: Synthetic Data Generation, Deep Convolutional Neural Network, Deep Learning, Computer Vision.

Abstract: One of the main limitations of the application of Deep Learning (DL) algorithms is when dealing with problems with small data. One workaround to this issue is the use of synthetic data generators. In this framework, we explore the benefits of synthetic data generation as a surrogate for the lack of large data when applying DL algorithms. In this paper, we propose a problem of learning to count the number of pedestrians using synthetic images as a substitute for real images. To this end, we introduce an algorithm to create synthetic images for being fed to a designed Deep Convolutional Neural Network (DCNN) to learn from. The model is capable of accurately counting the number of individuals in a real scene.

## 1 INTRODUCTION

Counting the number of objects in still images or video frames is a new approach towards dealing with detecting or learning objects which has been recently proffered in the literature (Rabaud and Belongie, 2006), (Kong et al., 2005). Previously, in order to count the objects of interest in an image or video, various object features needed to be designed, extracted or detected during the learning phase which restrict their usage in large-scale computer vision applications thus demanding more efficient solutions to alleviate, expedite and improve this process.

One of the recent and commonly used methods to facilitate feature detection process is the application of Deep Convolutional Neural Networks (DCNN) (Krizhevsky et al., 2012), (LeCun and Bengio, 2005), (Szegedy et al., 2015). One of the promises of DCNN is replacing handcrafted features with efficient algorithms for feature learning and hierarchical feature extraction (Song and Lee, 2013). DCNNs have been claimed and practically proven to achieve the most assuring performance in different vision benchmark problems concerning feature detection and classification (Ciregan et al., 2012), (Szegedy et al., 2015).

Although access to fast computers and vast amounts of data has enabled the advances of deep learning algorithms such as DCNN in solving many problems that were not solvable using classic AI, they have limitations. For instance, they do not perform well when there is limited data (Griffin et al., 2007). This constrain restricts the application of DL methods in various areas including computer vision where they have shown promising performances. As one solution to tackle this issue, we introduce a synthetic data generator algorithm to create images highly-representative of the real images.

In this article, we tackle a crowd counting problem by the means of synthetic images and deep convolutional neural network. We generate a set of highly realistic, synthetically generated images to be fed to a proposed convolution-based deep architecture. DCNN are well-suited for learning object features from the scratch and in a hierarchical approach. The proposed architecture consist of convolutional network to capture discriminative information about the object we are willing to count, following by fully connect layers where we count the multiplicity of object of interest. Figure 1 illustrates the proposal at a glance. The input instances contain a random set of pedestrians in a walkway. As it's shown in below, our goal is to learn to count the number of people in synthetic images and thereby, accurately predict the number of pedestrians in similar but real images.

Our contributions are as follows: We introduce a synthetic image generation algorithm in order to substitute the lacking training data in a fully supervised learning problem casted as learning to count the number of pedestrians in a walkway. Moreover, we propose a DCNN capable of learning pedestrians' features. Then, we validate our approach in a similar but real scenario. We test our proposed model which has been trained on synthetic images, on real images to see if synthetic data generation can be incorporated

Ekbatani, H., Pujol, O. and Segui, S.
Synthetic Data Generation for Deep Learning in Counting Pedestrians.
DOI: 10.5220/0006119203180323
In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pages 318-323
ISBN: 978-989-758-222-6

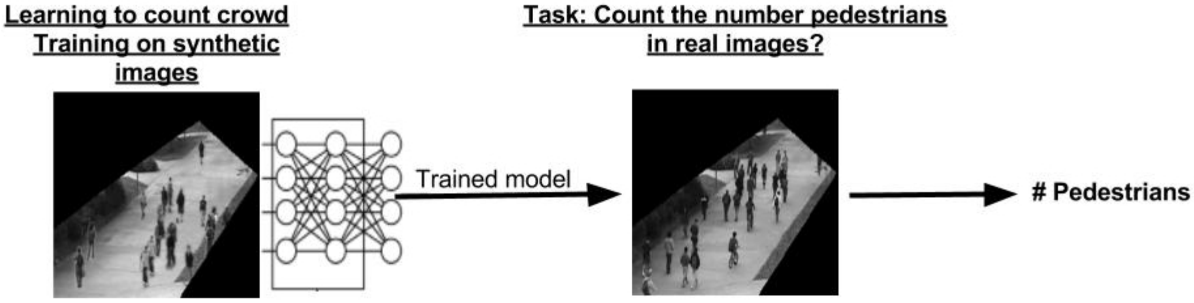
Learning to count crowd
Training on synthetic images
Task: Count the number pedestrians in real images?
Trained model
# Pedestrians

Figure 1: A schematic of our proposal. In this paper, we show that by creating realistic synthetic images, we are able to train a DCNN that is able to count the number of pedestrians in similar but real images.

as a surrogate for replacing small training sets when applying deep architectures.

## 2 BACKGROUND AND RELATED WORKS

### 2.1 Synthetic Data Generation

The main purpose of generating synthetic datasets has been to protect the privacy and confidentiality of the actual data (Phua et al., 2010), (Yao et al., 2013), since it does not hold any personal information and cannot be traced back by any individual. Problems such as fraud detection (Phua et al., 2010), or health care (Yao et al., 2013), are normally tackled by the use of synthetic data. However, most of the previously mentioned approaches towards synthetic data generation would not be applicable when it comes to synthetic image generation. This is due to the fact that standard methods such as Probability Density Function (PDF) or Interpolation operate element-wise. The need for generating and synthesizing images using object-wise operations led researchers to the use image processing tools for creating synthetic images to tackle vision problems.

In computer vision, usage of synthetic images has a longstanding history, as in 2000, Cappelli et al. in (Cappelli et al., 2000) presented an approach to synthetic fingerprint generation on the basis of some mathematical models that describe the main features of real finger prints. More recently, after the success of deep convolutional neural networks in various vision tasks concerning object detection or classification, generation and use of synthetic datasets has been frequently considered. For example, in (Eggert et al., 2015), synthetic images are generated to be fed to a DCNN in order to learn how to detect company logo in the absence of a large training set.

Moreover, as one of the most recent approaches, Segui et al. in (Seguí et al., 2015) proposed synthetic data generation to counter lack of data issue for learning to count the number of objects in images using deep convolutional neural networks. In their work, they took advantage of existent unlabeled and labeled datasets to generate synthetic images representative of the actual images. The authors introduce two counting problems, counting number of even-digits in images, and counting the amount of pedestrians in a walkway.

### 2.2 Crowd Counting

Learning to count the objects of interest in an image can be approached from two different perspectives: either training an object detector, or training an object counter. In the field of object detection, numerous works have been previously proposed (Kong et al., 2005), (Marana et al., 1998). Furthermore, Wu and Nevatia in (Wu and Nevatia, 2005) proposed edgelet features (an edgelet is a short segment of line or curve) as a new type of silhouette-oriented features to deal with the problem of detecting individuals in crowded still images.

As a similar line of work in the course of object counting and more specifically crowd counting, in (Leibe et al., 2007) and (Rabaud and Belongie, 2006), different object tracking approaches were taken to detect and count moving objects in the scene. However, most of object tracking approaches met with skepticism by society, given the perception of infringing individuals' privacy rights.

More recently, in (Chan et al., 2008), Chan et al. presented a novel approach with no explicit object segmentation or tracking to estimate the number of people moving in each direction (towards and away from camera) in a privacy-preserving manner.

On the other hand, in case of feature learning, Segui et al. in (Seguí et al., 2015) proposed a novel approach for counting objects representations using deep object features. In their work, objects' features

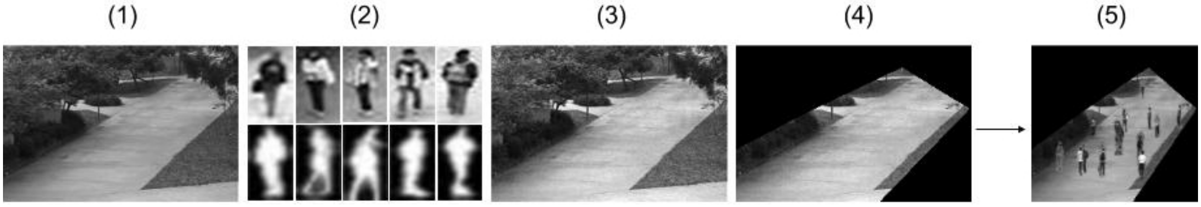

Figure 2: An illustration of image generation process at different steps.

are learned by a counting DCNN and are used to understand the underlying representation. Contrary to the previous approaches, their proposal is the first one where counting problem is handled by learning deep features. Additionally, no hints on the object of interest was given besides its' occurrence multiplicity.

# 3 SYNTHETIC IMAGE GENERATION

The main hypothesis of this work is that synthetic data generation algorithms can be used as a workaround for problems with no or little training sets. On this course, we propose an algorithm for creating highly realistic synthetic images of pedestrians in a walkway. We used UCSD unlabeled Anomaly detection dataset of pedestrians collected by Chan et al. and used in (Mahadevan et al., 2010) and (Chan et al., 2009). UCSD Anomaly detection dataset contains clips of groups of people walking towards and away from the camera, and consists of 34 training video samples and 36 testing video samples. Each video has 200 frames of each 238 × 158 pixels.

## 3.1 Image Generation

In our dataset, we employed all 70 training and testing video samples to generate the synthetic pedestrian dataset. We constrained each image by having up to 29 pedestrians in the walkway. The process of generating the data includes the following steps while figure 2 illustrates this process.

1. **Background Extraction.** Firstly, we simply subtract the background from each video frame and from there, we extract the median backgrounds of each video (in total, 70 different backgrounds).
2. **Pedestrian Extraction.** Subtracting each image from the mean background, we are able to label the connected regions (each individual in case of our images) using morphological labeling methods.
3. **Background Generation.** In this step, we try to make the backgrounds of images as realistic as possible by:
   - making a sparse combination of median backgrounds.
   - changing the global illumination of the images randomly.
   - adding some random Gaussian noise to the backgrounds.
4. **Region Of Interest (ROI).** Then, for training and comparison purposes, images are masked with a filter of Region Of Interest (ROI).
5. **Creating Synthetic Images.** Afterwards, pedestrians are added to the masked background in a way that the center of each person is placed inside white area of the mask. Finally images are normalized (between 0 and 255) and resized to 158 × 158 in order to be fed to convolution layers.

## 3.2 Image Improvement

Although we managed to successfully create synthetic images of people in the street, the generated images were still quite distinguishable from the real dataset. Thus, in order to make images as highly realistic as possible, we improved the dataset as explained underneath. Figure 3 depicts this procedure.

1. **Remove Non-pedestrians.** Amongst the extracted pedestrians, there were some non-pedestrians with objects instead of pedestrians, and yet others with more than one person. Therefore, we manually removed these outliers. After this edition, we ended with 426 samples of people.
2. **Lack of Pedestrians.** For the sake of generalization, we needed a decent variety of pedestrians in the images to train with. For this purpose, we created 2 versions of current pedestrians list, each darkened by the factor of 20% from each other.
3. **Halos Around the Pedestrians.** Due to lack of accuracy of the region measuring method, a fine layer of the background that pedestrians were extracted from, still remained around the pedestri-

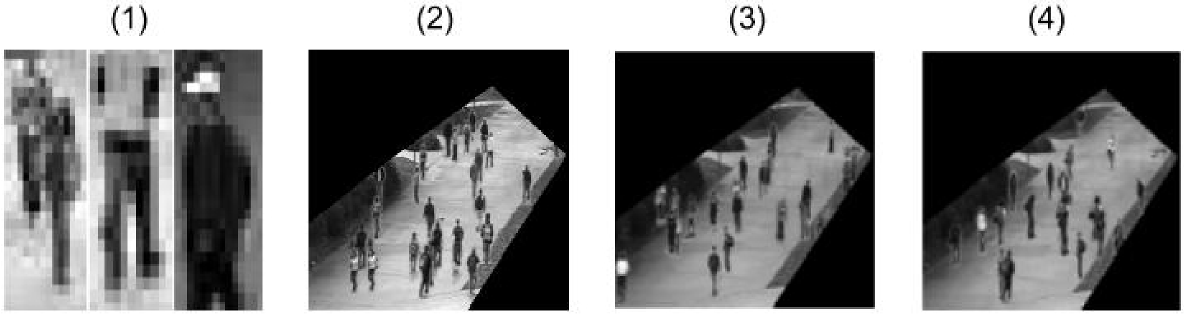

Figure 3: An illustration of each step of image improvement process.

ans. In the created images, depending on where the person was placed, these thin layers appeared like a halo around the person. We used *morphological erosion* on pedestrians' masks and also *Poisson image editing* to remove the halos.

4. **Image Perspective.** Finally, Since pedestrians of different sizes were put randomly in the images, we considered peoples tallness perspective in the images. Humans height almost follows a Gaussian distribution (Subramanian et al., 2011). Therefore, with respect to (Subramanian et al., 2011), we mapped individuals heights with the length of the walkway in the image, considering a Gaussian noise with mean $\mu = 0$ and $\sigma = 3.5$.

## 4 EXPERIMENTS AND RESULTS

For learning to count the number of pedestrians in a walkway, we synthetically generated a set of 1 million images of size 158 × 158 with up to 29 pedestrians in each image. Maximum overlapping was considered in the creation of the images. We divided this dataset into a training set of 800k images and 200k images for validation set. To test our model, we used UCSD crowd counting dataset with 3375 manually labeled images of pedestrians. The selected UCSD images contain from 11 to 29 pedestrians in each image.

We designed a seven layers architecture DCNN with four convolutional layers and three fully connected layers. The architecture is shown in Table 1.

Table 1: Proposed DCNN for counting pedestrians.

| Convolutions | Fully connects |
|---|---|
| $10 \times 15 \times 15$ & x2 pooling | 128 |
| $10 \times 11 \times 11$ & x2 pooling | 64 |
| $20 \times 9 \times 9$ | 1 |
| $20 \times 5 \times 5$ | |

The algorithm is trained using the Caffe package[11] on a GPU NVIDIA Tesla K40. The network has been set to 400,000 iterations. The output layer is configured as a classification problem.

On the validation set, the performance of the model is **0.70** mean absolute error and **0.94** mean squared error. This results improve the achieved results in a similar experiment done by (Seguí et al., 2015) (the comparison is shown in table 2). On the other hand, on the real test set, we obtained **1.38** mean absolute error and **3.61** mean squared error which closely follow the results in (Chan et al., 2008) which was obtained by hand-crafting highly specialized image features that are dependent on the object class. This comparison is depicted in table 3 The confusion matrix regarding the model performance is illustrated in figure 4. As you may notice, due to the inevitable differences between real and synthetic samples, the model mostly over-predicts. Moreover, as the number of pedestrians increases in the images, the prediction accuracy of the model decreases.

Table 2: Performance comparison on the synthetic data between our proposal and related work in (Seguí et al., 2015).

| Experiments | MSE | MAE |
|---|---|---|
| Our approach (29 peds) | 0.942 | 0.707 |
| (Seguí et al., 2015) (25 peds) | 1.12 | 0.74 |

Table 3: Performance comparison on the real data between our proposal and related work in (Chan et al., 2008).

| Experiments | MSE | MAE |
|---|---|---|
| Proposed method | 3.61 | 1.38 |
| (Chan et al., 2008) approach | 2.73 | 1.24 |

As you may observe in table 2, in case of synthetic images, although our images contain more pedestrians, our results beat the previous approach in (Seguí et al., 2015). This proves the improvement we made in synthetic data generation process and the designed deep architecture.

Respectively, in case of real images, although we could not improve the work done in (Chan et al., 2008), our results follows their results closely. We

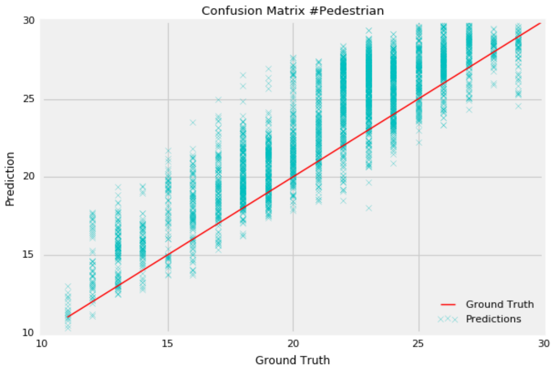

Figure 4: Confusion matrix regarding the model performance on the real test set. The starting point of the graph is 11 since the minimum amount of pedestrians in the real test set is 11.

should mention that Chan et.al experiment in (Chan et al., 2008) was done by hand-crafting highly specialized features and exhaustive labeling. This results approve the suitability of synthetic data as a surrogate for the small real data when using DCNN.

## 5 CONCLUSIONS

In this paper we explore the benefits of synthetic data generation for the application of deep convolutional neural networks for a crowd counting problem with small training set. We propose an algorithm for creating a highly realistic synthetic dataset of pedestrians in a walkway to train the proposed DCNN with. Moreover, we provide a system trained with synthetic images capable of predicting the number of pedestrians in an image to a satisfactory extent. The obtained results suggest the incorporation of synthetic data as a well-suited surrogate for the missing real along with alleviating required exhaustive labeling.

There are still many open questions to be addressed such as, when and to what extent synthetic images are applicable as a substitute to solve real world problems. which is the best network architecture for counting the crowd?

## ACKNOWLEDGEMENTS

This work has been partially funded by the Spanish MINECO Grants TIN2013-43478-P and TIN2012-38187- C03. We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Tesla K40 GPU used for this research.

## REFERENCES

Cappelli, R., Erol, A., Maio, D., and Maltoni, D. (2000). Synthetic fingerprint-image generation. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*. IEEE.

Chan, A. B., Liang, Z.-S. J., and Vasconcelos, N. (2008). Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE.

Chan, A. B., Morrow, M., and Vasconcelos, N. (2009). Analysis of crowded scenes using holistic properties. In *Performance Evaluation of Tracking and Surveillance workshop at CVPR*.

Ciregan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE.

Eggert, C., Winschel, A., and Lienhart, R. (2015). On the benefit of synthetic data for company logo detection. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM.

Griffin, G., Holub, A., and Perona, P. (2007). Caltech-256 object category dataset. California Institute of Technology.

Kong, D., Gray, D., and Tao, H. (2005). Counting pedestrians in crowds using viewpoint invariant training. In *BMVC*. Citeseer.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.

LeCun, Y. and Bengio, Y. (2005). Convolutional networks for images, speech, and time series. In *BMVC*. Citeseer.

Leibe, B., Schindler, K., and Van Gool, L. (2007). Coupled detection and trajectory estimation for multi-object tracking. In *2007 IEEE 11th International Conference on Computer Vision*. IEEE.

Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In *CVPR*.

Marana, A., Costa, L. d. F., Lotufo, R., and Velastin, S. (1998). On the efficacy of texture analysis for crowd monitoring. In *Computer Graphics, Image Processing, and Vision, 1998. Proceedings. SIBGRAPI'98. International Symposium on*. IEEE.

Phua, C., Lee, V., Smith, K., and Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. In *arXiv preprint arXiv:1009.6119*.

Rabaud, V. and Belongie (2006). Counting crowded moving objects. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE.

Seguí, S., Pujol, O., and Vitria, J. (2015). Learning to count with deep object features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Song, H. A. and Lee, S.-Y. (2013). Hierarchical representation using nmf. In *International Conference on Neural Information Processing*. Springer.

Subramanian, S., Özaltin, E., and Finlay, J. E. (2011). Height of nations: a socioeconomic analysis of cohort differences and patterns among women in 54 low-to middle-income countries. In *PLoS One*. Public Library of Science.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wu, B. and Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. IEEE.

Yao, W., Basu, S., Wei-Nchih, L., and Singhal, S. (2013). Synthetic healthcare data generation. Google Patents.