Investigating Synthetic Data Sets for Crowd Counting in Cross-scene Scenarios

Rita Delussu, Lorenzo Putzu and Giorgio Fumera
University of Cagliari, Piazza D'Armi, Cagliari, Italy
Department of Electrical and Electronic Engineering, Piazza D'armi, 09123 Cagliari, Italy

Keywords: Cross-scene, Crowd Analysis, Crowd Density Estimation, Synthetic Data Sets, Texture Features, Regression.

Abstract: Crowd counting and density estimation are crucial functionalities in intelligent video surveillance systems but are also very challenging computer vision tasks in scenarios characterised by dense crowds, due to scale and perspective variations, overlapping and occlusions. Regression-based crowd counting models are used for dense crowd scenes, where pedestrian detection is infeasible. We focus on real-world, cross-scene application scenarios where no manually annotated images of the target scene are available for training regression models, but only images with different backgrounds and camera views can be used (e.g., from publicly available data sets), which can lead to low accuracy. To overcome this issue, we propose to build the training set using *synthetic* images of the target scene, which can be automatically annotated with no manual effort. This work provides a preliminary empirical evaluation of the effectiveness of the above solution. To this aim, we carry out experiments using real data sets as the target scenes (testing set) and using different kinds of synthetically generated crowd images of the target scenes as training data. Our results show that synthetic training images can be effective, provided that also their background, beside their perspective, closely reproduces the one of the target scene.

1 INTRODUCTION

The use of computer vision tools to automatise crowd monitoring and analysis tasks, or to support human operators involved in such tasks, is becoming increasingly relevant in many applications, such as crowd behaviour analysis and intelligent video surveillance, given the nowadays pervasive deployment of CCTV systems. One prominent example is the use of CCTV systems by Law Enforcement Agencies to monitor and guarantee the security of mass gathering events, which is one of the issues addressed by the EU H2020 LETSCROWD project [1] our research group is working on. In particular, crowd counting and density estimation are potentially very useful functionalities of crowd monitoring systems. These are however very challenging tasks in practical applications due to issues such as illumination changes, severe occlusions due to objects in the scene or by other people (see Fig. 1), and distortions caused by the camera view, which make the size and perspective of people in the scene change considerably according to the distance from the camera. In this work, we focus on the most challenging scenario of dense crowd scenes (see Fig. 2), where crowd counting and density estimation cannot rely on pedestrian detection. For dense crowd scenes, the so-called counting by regression approach is widely used in the literature (Loy et al., 2013). It consists of estimating the number of people using a regression model trained on low-level image

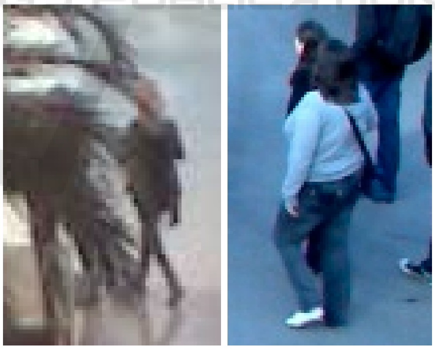

Figure 1: Example of a static (left) and dynamic (right) occlusion, where a person is almost totally occluded by a palm and by another person, respectively. (Images taken from the Mall and PETS2009 data sets – see Sect- 3.3).

[1] https://letscrowd.eu/

Delussu, R., Putzu, L. and Fumera, G.
Investigating Synthetic Data Sets for Crowd Counting in Cross-scene Scenarios.
DOI: 10.5220/0008981803650372
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 365-372
ISBN: 978-989-758-402-2; ISSN: 2184-4321

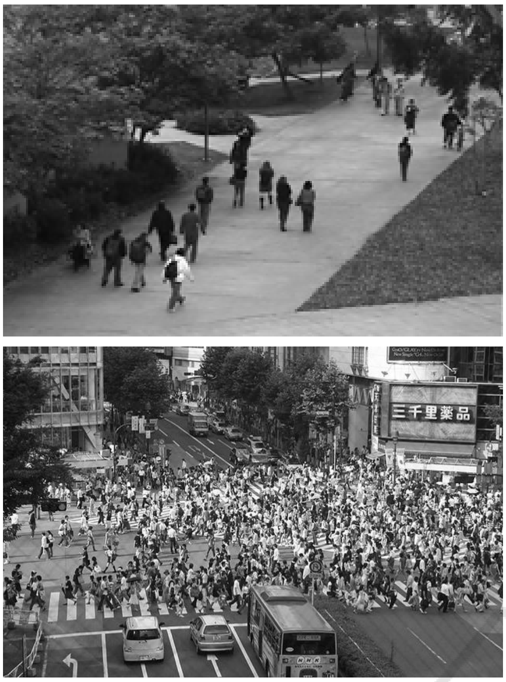

Figure 2: Examples of a non-crowded scene (top) and of a crowded scene (bottom). (Images taken from the UCSD and ShanghaiTech data sets – see Sect- 3.3).

features, using images of a crowd manually annotated with the number of people in the scene. Recently, several regression-based methods that use Convolutional Neural Networks (CNNs) have been proposed (Sindagi and Patel, 2017a); most of them directly estimate the density map and then obtain the people count from it. However, CNNs require massive amounts of training data, which may not be available in real applications. Moreover, using the density map as the ground truth requires the position of each pedestrian in the scene to be manually annotated (Liu et al., 2019), which requires a much higher effort than providing only the people count. Furthermore, CNN-based approaches may require several seconds to compute the density map for a single video frame, which makes them not yet suitable for real-time applications. For all the reasons above in this work, we focus on earlier regression-based approaches that do not use CNNs.

Two issues that affect the performance of regression-based methods are the fact that the features typically used are significantly affected by the image background, and that in some application scenarios it may be unfeasible to collect and manually annotate crowd images of the *target* scene to build a representative training set. In such a case one faces a challenging cross-scene scenario characterised by training images exhibiting different perspective (camera views) and background from the target scene where the crowd counting system has to be deployed: as a consequence its accuracy can be very low. This is the application scenario we focus on in this work.

Inspired by some recent works that used *synthetic* training images to improve the performance of crowd counting approaches based on pedestrian or body part detection (Courty et al., 2014; Schröder et al., 2018), as well as in other computer vision tasks, in this work we propose to address the issues of regression-based crowd counting methods mentioned above by building a training set of *synthetic* images of the target scene. This would allow, e.g., to create scene-specific regression models on the fly for new camera installations. This work aims to empirically investigate whether and to what extent the above solution can improve the cross-scene effectiveness of existing regression-based crowd counting methods in dense crowd scenes. To this aim, we carry out experiments using three real, benchmark data sets of dense crowd images as the target scenes (testing set) and using different kinds of synthetically generated crowd images of the same target scenes as training data.

The remainder of the manuscript is structured as follows. Related works are first summarised in Sect. 2. Sect. 3 describes the features and the regression models considered, the real and the synthetic data sets used in our experiments, and the accuracy measures used for performance evaluation. The experimental set-up and the results are presented in Sect. 4. Sect. 5 summarises the conclusions of this work.

2 RELATED WORK

Several crowd counting approaches have been proposed so far. They can be divided into counting by detection, counting by clustering and counting by regression (Loy et al., 2013). Counting by detection is based on pedestrian detection from still images, either full-body (Leibe et al., 2005; Dollar et al., 2011) or body part detection (Lin et al., 2001; Tu et al., 2008). The latter aims at overcoming the presence of occlusions and attempts to locate head and shoulders (Ge and Collins, 2009), as these are among the most important and most visible cues for typical camera locations (Tu et al., 2008). However counting by detection is effective only under very limited occlusion, i.e., for non-crowded scenes (see Fig. 2, left), whereas it is unable to deal with dense crowds (see Fig. 2, right).

Counting by clustering is based on people tracking, and assumes that coherent feature trajectories can be grouped together to approximate the number of people; this approach is ineffective as well on dense crowd scenes (Loy et al., 2013).

The counting by regression approach, which we focus on in this work, aims at mapping from low-level image features to the number of people or to the density map of a scene by supervised training of a regression model. The earliest methods estimate the number of people using holistic scene descriptors such as foreground segment (Ma et al., 2004), edge (Kong et al., 2005), texture and gradient (Wu et al., 2006; Ojala et al., 2002), shape (Dong et al., 2007), intensity (Lempitsky and Zisserman, 2010) and motion (Benabbas et al., 2010). The main regression models used in such methods are Linear Regression, Random Forests and Support Vector Regression (Loy et al., 2013). More recently, a variety of methods based on CNNs, mostly regression-based, have been proposed (Sindagi and Patel, 2017a). Such methods estimate the density map directly and obtain the people count from it. Most of them exploit specifically-devised CNN architectures, although they often share several layers with generic architectures and then fine-tune them on a specific training set (Sindagi and Patel, 2017b).

As for all supervised methods, the effectiveness of regression-based crowd counting methods relies on a representative training set of crowd images manually annotated with the people count (or the density map). However, in some application scenarios, it may be unfeasible to collect and manually annotate crowd images to build a training set representative of the *target* scene where a crowd counting system has to be deployed. If a training set made up of images of *different* scenes (e.g., taken from publicly available, benchmark data sets), one faces a cross-scene scenario where the mismatch in image background and perspective (camera view) can severely affect the accuracy of the resulting model. We shall provide clear empirical evidence of this issue in Sect. 4. In fact, features typically used by regression-based methods are significantly affected by the image background. Existing background subtraction and segmentation approaches are not sufficient under frequent illumination changes or, even worse, when a background image is not available. Moreover, regression models are affected by perspective distortions (objects in the background appear smaller than foreground ones).

Several data sets of crowd images have been collected so far (see Sect. 3.3), which however exhibit several limitations: most of them contain only non-dense crowd scenes with limited occlusion, and are also relatively small in size. This can further affect the cross-scene performance of regression-based models. Domain adaptation methods have been proposed to mitigate the cross-scene issue (Change Loy et al., 2013), but they nevertheless require images of the target scene for fine-tuning, which is not feasible in the application scenario considered in this work.

In some computer vision tasks, including crowd counting by detection, the use of *synthetic* data sets built using computer graphics tools has been proposed to overcome the limits of data sets made up of real images. This solution can be potentially useful also for regression-based crowd counting in the above application scenario since it would allow to generate synthetic images of the target scene, and to automatically control every parameter of interest such as the number and location of pedestrians and the scene perspective, background and illumination. A similar solution has already been proposed in (Wang et al., 2019), but in the context of a domain adaptation method which requires crowd images of the target scene for fine-tuning, which is not feasible in the application scenario considered in this work.

3 EXPERIMENTAL SETTING

In this section, we describe the methods, the data sets and experimental setting used in this work. Since we focus on the counting by regression approach, we first describe the features and regression models, that are mainly based on the ones described in the surveys of (Loy et al., 2013; Ryan et al., 2015).

3.1 Feature Extraction

Several kinds of features have been proposed so far for regression-based crowd counting, and often different complementary features are combined together. We consider here segment and edge features, which are among the most common foreground ones, as well as the Grey-Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP) texture features. Foreground features can be obtained through background subtraction. Segment features (Ma et al., 2004) aim at capturing *global* properties of the image regions, such as area and perimeter; edge features (Kong et al., 2005) focus on complementary information about *local* image properties, such as the number of edge pixels and edge orientation. Texture features analyse the spatial relationships among image pixels (Loy et al., 2013) instead. In particular, GLCM is obtained in the form of a matrix by computing how often pairs of pixels with a certain value and in a specified spatial relationship occur. From the GLCM several global statistical information can be extracted (Haralick et al., 1973). The LBP descriptor is capable to characterise the local structure of the image, as it is rotation invariant and robust to grey-scale changing

(Ojala et al., 2002). All the mentioned feature sets have been combined together, using a simple concatenation, to create a single and stronger feature set.

3.2 Regression Models

Regression-based methods can be divided into global and local (Loy et al., 2013). Global methods learn a single regression function for the whole image, whereas local methods subdivide an image into patches and estimate the total people count by performing regression on each patch. Local methods can handle efficiently different kinds of scenes in which the density is not the same over the whole image but can be too complex for real-world applications. For this reason, we focus on global methods.

Several global regression models have been proposed in the literature (Loy et al., 2013). The simplest one is linear regression, which estimates the people count finding a linear relationship with a D-dimensional feature vector that describes the image. Since the feature space size D can be very high, to address the issue of feature collinearity partial least squares (PLS) regression has been used (Abdi, 2010). Non-linear models have also been proposed for crowd counting, such as Gaussian process regression (GPR). However, its processing cost at the prediction phase is too high and therefore, not suitable for real-time applications. For this reason, we do not consider it in this work, preferring others non-linear models, such as the Random Forests (RF) (Criminisi et al., 2012), which has the advantage of being scalable and less sensitive to parameter selection, and the Support vector regression (SVR) (Basak et al., 2007) with a radial basis function (RBF) kernel, which is one of the most popular regression models.

3.3 Real Data Sets

As stated above, in this work we focus on crowd counting systems to be deployed on a *specific* target scene (i.e., camera view), with a *dense* crowd. To our knowledge, only three publicly available data sets can be considered representative of dense crowd scenarios (Zhang et al., 2016; Zhang et al., 2019; Sindagi and Patel, 2017a), namely ShanghaiTech, UCF-QNRF and World Expo Shanghai 2010. However, they do not contain images (in a significant number) from a *same* scene. On the contrary, to evaluate the effectiveness of a crowd counting system data sets characterised by the whole *videos* (or at least a significant number of frames) from a *same* scene with a *dense* crowd are required. Unfortunately, there are very few publicly available data sets manually annotated with the crowd count that exhibit all the above features. For this reason, we used only three real data sets of non-dense crowd scenes (less than 60 persons per image), that are Mall, UCSD and PETS. **[Mall]** has 2000 frames collected using a publicly accessible camera in a shopping mall (Loy et al., 2013). It contains several crowd densities from 13 to 53 people per frame (on average 31) for a total of 62,325 pedestrians. This is a challenging data set with severe perspective distortion and several occlusions caused by static objects or by other people. **UCSD** contains a total of 70 videos acquired from a low-resolution camera (frames of size 238×158) installed in a pedestrian walkway at the UCSD campus (Chan et al., 2008). It contains a total of 49,885 pedestrians, with an average number of people per frame of around 25. For these data sets, we used the same setting as in (Loy et al., 2013), where 2000 frames were extracted from the original data set. **[PETS2009]** was built to test several visual surveillance tasks (Ferryman and Shahrokni, 2009), including people counting (S1 part). It is a multi-view data set, and each sequence has been acquired with a different camera, but the frames belonging to the same camera view can be used to create some single-scene data set. To this aim, we used only the images from the first camera as a single-scene data set, using the ground truth provided in (Zhang and Chan, 2019). Fig. 3 shows some examples of frames from all the above data sets.

3.4 Synthetic Data Sets

Building and evaluating crowd counting models in real-world application scenarios is a very complex task, especially when no annotated images of the target scene are available to train regression models. Synthetic images can be very useful to this aim, since the exact number of pedestrians shown in such images is exactly known and no manual counting is required. Synthetic data sets are therefore a potentially useful solution to build the training set of a regression model. To this aim, in this work we evaluated the effectiveness of using as a training set synthetic crowd images which reproduce the target scene (the background, the perspective or both) of the testing set, where the latter is obtained from a real data set.

We collected three different data sets of synthetic images. The first two data sets have been created using a commercial crowd modelling and simulation software developed by a member of the LETSCROWD consortium, [2] which is based on the computer graphic engine Unity [3]. This software al-

[2] Crowd Dynamics, https://www.crowddynamics.com/

[3] https://unity.com/

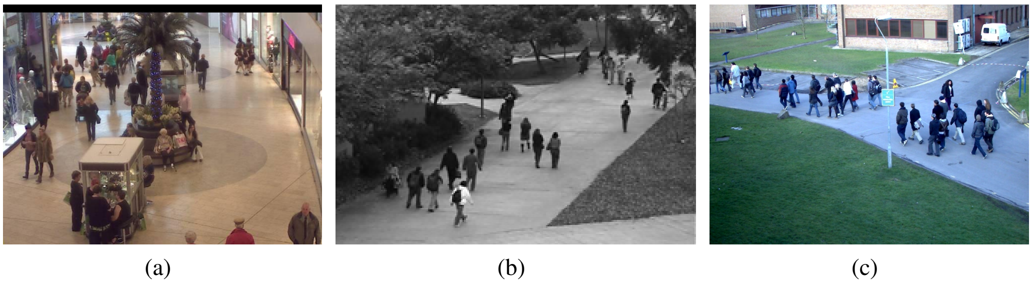

Figure 3: Example of frames from real data sets: (a) Mall, (b) UCSD, (c) PETS 2009.

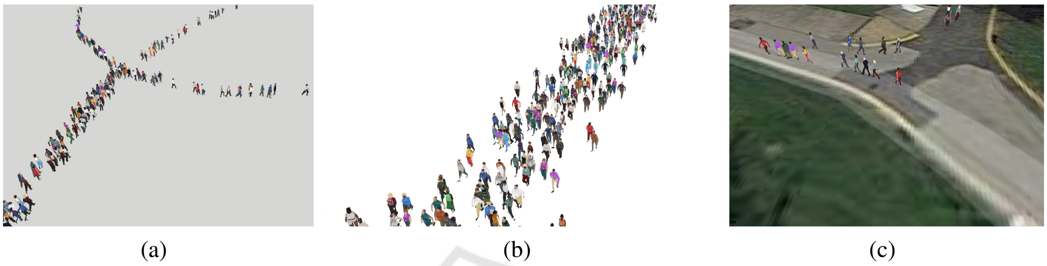

Figure 4: Example of frames from our CG-sets: (a), (b) uniform background; (c) PETS2009-like background.

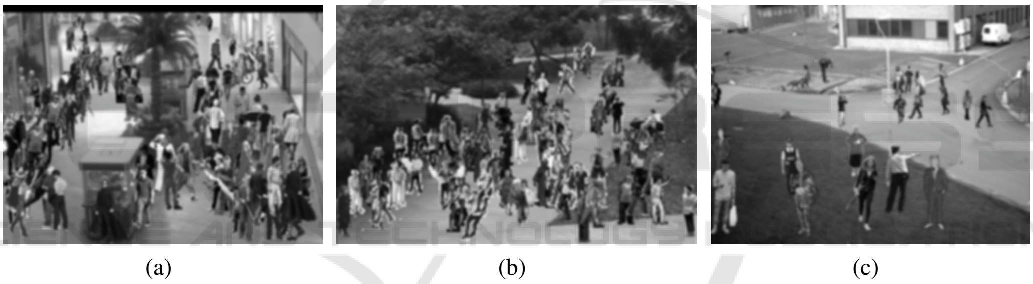

Figure 5: Examples of frames from our PNG-sets, using background from the real data sets: (a) Mall, (b) UCSD, (c) PETS 2009. (Images are slightly blurred in this figure for privacy reasons.)

lows to generate synthetic crowd scenes by choosing a background image, the number of pedestrians and their paths. In the first data set, we reproduced the camera views and perspectives of the three real data sets, but not the corresponding people flow and crowd size, nor the background. We collected in total 15000 frames from videos of pedestrians walking in different directions on a uniform background; the number of pedestrians in each frame ranges from 1 to 1000.

Since the features used to train global regression models are significantly affected by the image background, in the second set of synthetic images, we tried to simulate also the background of the target, real testing scenes. However, this turned out to be much more complex, and therefore we only reproduced the single view extracted from the PETS2009 real data set. Fig. 4 shows some frames of the first two synthetic data sets above obtained by a computer graphic tool, which will be denoted from now on by CG-sets.

In the third data set of synthetic images, we simulated the real testing scene by superimposing pedestrian images to a *real* background image of that scene, to increase the realism of synthetic images. First, for each real (testing) data set we picked an image where no pedestrians appeared (if any), or obtained a background image by background subtraction; then we defined the region of interest (ROI) and the perspective map. The perspective map is necessary to estimate the pedestrian scales (Zhang et al., 2015), and is computed by linear interpolation from the measurements of several pedestrians randomly selected in the images, assuming that all the adult pedestrians present a standard height. This map allows one to easily compute the height in pixels of a pedestrian in each location of the ROI. Then we collected from the web a set of images with a transparent background (in PNG format) showing a single person. The synthetic images were finally obtained by superimposing a given number of randomly selected pedestrian images over the background image at randomly chosen locations in the ROI, by resizing pedestrian images according to the perspective map. Overlapping between pedestrians was controlled by drawing first the one farthest from the camera. For each target (test) scene, we built 1000 synthetic images containing a number of pedestrians ranging from 1 to 100. Fig. 5 shows some examples of these images, that from now on will be denoted by PNG-sets. We point out that this latter kind of synthetic images may not present a realistic perspective distortion nor real human poses. Nevertheless, such images respect the scale and the camera view, which are the most relevant features for the considered task; this is also a straightforward procedure to create synthetic images that reproduce a given target scene. In particular, it also allows to automatically annotate each image with the number of pedestrians inside it.

3.5 Performance Measures

We evaluated crowd counting accuracy using two common metrics that are defined over a single image: the absolute error (AE) and the root squared error (RSE). We shall report their average values computed across all testing images of a given target scene, i.e., the mean absolute error (MAE) and the root mean squared error (RMSE):

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\eta_i - \hat{\eta}_i| \, , \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\eta_i - \hat{\eta}_i)^2} \, , \tag{2}$$

where N is the total number of testing images, η_i is the ground truth (pedestrian number) and $\hat{\eta}_i$ is the estimated pedestrian number for the i-th frame. The MAE computes the average absolute difference between the actual and the estimated count. The second RMSE, as a result of the squaring of each difference, penalises large errors more heavily than small ones, instead. We point out that in many recent works (Liu et al., 2019; Wang et al., 2019; Ryan et al., 2015) the metric RMSE was used, but it was called MSE; however, the MSE metric does not include the square root (Loy et al., 2013). We prefer to use the RMSE metric to be aligned with the most recent works in this area, and also because it is expressed in the same unit of measurement as MAE.

Table 1: Crowd counting accuracy (MAE and RMSE) in a same-scene scenario (training and testing images come from the same scene) using different regression models (RF, SVR and PLS), on each real data set (Mall, UCSD and PETS).

Data set	RF		SVR		PLS	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Mall	3.66	4.53	5.53	6.98	3.03	3.83
UCSD	2.85	3.45	6.79	8.35	2.32	2.82
PETS	7.51	9.49	8.66	10.34	5.46	7.1

4 EXPERIMENTS

We performed three different experiments whose goal is to verify if and to what extent using synthetic images of the target scene in the training set can provide an effective, scene-specific regression model for people counting tasks in dense crowd scenes. To create a baseline for comparison, in the first experiment we evaluated the performance on each of the three real data sets of Sect. 3.3 by using in the training and test set only *real* images from the *same* data set. To average the influence of training/testing data splits we performed a 10-fold cross-validation, where the training and test sets are made up respectively of 80% and 20% disjoint subset of images. The results are reported in Table 1. The second experiment simulates a real-world cross-scene application scenario, e.g., when a camera is installed in a new place and no annotated crowd images of the same target scene are available for training a scene-specific regression model; in this case the regression model has to be trained using a previously available data set which may be not representative of the target scene. To this aim we trained a regression model on training data belonging to one or two of the three *real* data sets, and tested it on the other data set, which is considered as the target scene. Our aim is to evaluate the amount of performance decrease with respect to the use of training data from the *same* target scene, as in the first experiment above. The results are reported in Table 2. As it can be expected, a comparison between tables 1 and 2 shows that the performance on the target scene always decrease when training data come from different scenes, which can be due to the difference in the background and in the perspective. We also point out that using the perspective correction described in (Loy et al., 2013) we observed a further performance decrease: this may be because the perspective maps used are not invariant to scale.

The third experiment aims at evaluating whether using *synthetic* training images of the target scene improves the crowd counting accuracy with respect to

Table 2: Crowd counting accuracy (MAE and RMSE) in a cross-scene scenario (training and testing images come from different scenes) using different regression models (RF, SVR and PLS). Each group of three rows corresponds to a single real data set used as the target scene (Mall, UCSD and PETS); the rows in each group correspond to the other two data sets used for training, either alone or together.

	Train Set	RF		SVR		PLS	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
Mall	UCSD	17.16	18.51	6.39	8.09	15.57	16.75
	PETS	16.26	17.21	8.87	10.71	19.38	20.08
	UCSD-PETS	18.32	19.47	6.89	8.65	14.49	15.83
UCSD	Mall	10.92	12.49	9.44	10.62	10.53	13.05
	PETS	12.39	14.02	20.97	21.42	50.76	51.37
	Mall-PETS	13.08	14.86	19.97	20.23	19.78	19.99
PETS	Mall	18.26	20.97	13.27	15.51	23.84	26.29
	UCSD	10.11	12.79	10.75	12.46	10.36	13.11
	Mall-UCSD	13.77	15.82	12.18	15.08	16.51	18.52

Table 3: Crowd counting accuracy attained by training the regression models RF, SVR and PLS on synthetic images reproducing the target (testing) scene. Synthetic images come from the CG-set: 'Blank-CG' and 'PETS-CG' denote respectively images with no background, and images which reproduce the PETS data set background.

Test	Train	RF		SVR		PLS	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
Mall	Blank-CG	5.74	7.16	5.52	6.97	5.62	7.0
UCSD	Blank-CG	5.9	7.47	6.99	8.37	7.24	12.17
PETS	Blank-CG	9.25	11.66	9.31	10.9	7.56	9.39
PETS	PETS-CG	7.39	9.28	8.55	10.31	7.62	9.24

the ones attained by using *real* training images from *different* scenes. The set-up is similar to the second experiment, with the only difference that, for each real data set used as the target scene (testing set) each regression model is now trained using one of the data sets of synthetic images (see Sect. 3.4) that reproduce the same target scene. The results are reported in Table 3 for CG-set images and in Table 4 for PNG-set images. As expected, the crowd counting performances on the target scene are better when training data come from synthetic images, instead of from real images of a different scene, as in Table 2. This mainly depends on the background and perspective that could differ a lot between two real scenes, while using the synthetic images we can reproduce, with a certain accuracy, both features. Nevertheless, the performances of the regression models exploiting synthetic images are still lower than the ones obtained from models exploiting real images of the same scene (reported in Table 1), even if the gap in performances become smaller with the presence of more realistic (accurate) images. This is also confirmed by comparing the results in Table 3 and in Table 4, where the regression models trained with the PNG-set images demonstrated better performances than the ones obtained with the CG-set images. In general, the presence of the real scene background in the synthetic images is crucial to create strong regression models. Indeed, even using the CG-set which reproduce the PETS data set background, the performances increase compared to the images presenting a blank background. This shows that also fully synthetic images could be effective to some extent, but they would require a background and a perspective view closer to the images of the target scene.

Table 4: Crowd counting accuracy attained by training the regression models RF, SVR and PLS on synthetic (PNG-set) images reproducing the target (testing) scene.

Test	Train	RF		SVR		PLS	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
Mall	Mall-PNG	4.73	5.84	5.99	7.63	5.65	7.14
UCSD	UCSD-PNG	4.08	4.95	11.04	13.35	2.76	3.5
PETS	PETS-PNG	7.42	9.07	14.77	17.94	8.11	10.31

5 CONCLUSIONS

In this work, we considered a challenging, cross-scene application scenario for crowd counting in dense crowd scenes, where no manually annotated images of the target scene are available for training a regression-based model. To improve the accuracy of regression models under this scenario, we proposed to build the training set using synthetic images of the target crowd scene, characterised by the same perspective (view) and possibly by the same background. We empirically evaluated the effectiveness of this solution using several real data sets of crowd images as the target scenes (testing set) and different kinds of synthetically generated images of the target scene as the training set. Preliminary results provide evidence that using synthetic images can be an effective solution, provided that they closely reproduce the background and the camera perspective of the target scene.

The use of synthetic images may also be useful in the same application scenario considered here for regression-based methods that use CNNs, which have not been considered in this work due to their still too high processing time during inference. Accordingly, our ongoing efforts are devoted to investigating the performance of CNN-based crowd counting methods using synthetic images for training or fine-tuning.

ACKNOWLEDGEMENTS

This work has been partially supported by the LETSCROWD project, funded by the European Union Horizon 2020 research and innovation programme under grant agreement No 740466, and by the project ODIS - Optimisation of DIstributed systems in the Smart-city and smart-grid settings, CUP:F72F16003170002, funded by Fondazione di Sardegna. The authors also thank the LETSCROWD partner Crowd Dynamics for providing some of the synthetic data sets.

REFERENCES

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (pls regression). *Wiley interdisciplinary reviews: computational statistics*, 2(1):97–106.

Basak, D., Pal, S., and Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10):203–224.

Benabbas, Y., Ihaddadene, N., Yahiaoui, T., Urruty, T., and Djeraba, C. (2010). Spatio-temporal optical flow analysis for people counting. In *AVSS*, pages 212–217.

Chan, A. B., Liang, Z.-S. J., and Vasconcelos, N. (2008). Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, pages 1–7.

Change Loy, C., Gong, S., and Xiang, T. (2013). From semi-supervised to transfer counting of crowds. In *ICCV*, pages 2256–2263.

Courty, N., Allain, P., Creusot, C., and Corpetti, T. (2014). Using the agoraset dataset: Assessing for the quality of crowd video analysis methods. *Pattern Recognition Letters*, 44:161–170.

Criminisi, A., Shotton, J., Konukoglu, E., et al. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2–3):81–227.

Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. on PAMI*, 34(4):743–761.

Dong, L., Parameswaran, V., Ramesh, V., and Zoghlami, I. (2007). Fast crowd segmentation using shape indexing. In *ICCV*, pages 1–8.

Ferryman, J. and Shahrokni, A. (2009). Pets2009: Dataset and challenge. In *PETS*, pages 1–6.

Ge, W. and Collins, R. T. (2009). Marked point processes for crowd counting. In *CVPR*, pages 2913–2920.

Haralick, R. M., Shanmugam, K., et al. (1973). Textural features for image classification. *IEEE Trans. on systems, man, and cybernetics*, SMC-3(6):610–621.

Kong, D., Gray, D., and Tao, H. (2005). Counting pedestrians in crowds using viewpoint invariant training. In *BMVC*, pages 1–10.

Leibe, B., Seemann, E., and Schiele, B. (2005). Pedestrian detection in crowded scenes. In *CVPR*, pages 878–885.

Lempitsky, V. and Zisserman, A. (2010). Learning to count objects in images. In *Advances in neural information processing systems*, pages 1324–1332.

Lin, S.-F., Chen, J.-Y., and Chao, H.-X. (2001). Estimation of number of people in crowded scenes using perspective transformation. *IEEE Trans. on Systems, Man, and Cybernetics*, 31(6):645–654.

Liu, X., Van De Weijer, J., and Bagdanov, A. D. (2019). Exploiting unlabeled data in cnns by self-supervised learning to rank. *Trans. on PAMI*, 41(8):1862–1878.

Loy, C. C., Chen, K., Gong, S., and Xiang, T. (2013). Crowd counting and profiling: Methodology and evaluation. In *Modeling, simulation and visual analysis of crowds*, pages 347–382. Springer.

Ma, R., Li, L., Huang, W., and Tian, Q. (2004). On pixel count based crowd density estimation for visual surveillance. In *CIS*, pages 170–173.

Ojala, T., Pietikäinen, M., and Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on PAMI*, 24(7):971–987.

Ryan, D., Denman, S., Sridharan, S., and Fookes, C. (2015). An evaluation of crowd counting methods, features and regression models. *Computer Vision and Image Understanding*, 130:1–17.

Schröder, G., Senst, T., Bochinski, E., and Sikora, T. (2018). Optical flow dataset and benchmark for visual crowd analysis. In *AVSS*, pages 1–6.

Sindagi, V. and Patel, V. M. (2017a). A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, 107:3–16.

Sindagi, V. A. and Patel, V. M. (2017b). Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *AVSS*, pages 1–6.

Tu, P., Sebastian, T., Doretto, G., Krahnstoever, N., Rittscher, J., and Yu, T. (2008). Unified crowd segmentation. In *ECCV*, pages 691–704. Springer.

Wang, Q., Gao, J., Lin, W., and Yuan, Y. (2019). Learning from synthetic data for crowd counting in the wild. In *CVPR*, pages 8198–8207.

Wu, X., Liang, G., Lee, K. K., and Xu, Y. (2006). Crowd density estimation using texture analysis and learning. In *ROBIO*, pages 214–219.

Zhang, C., Li, H., Wang, X., and Yang, X. (2015). Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, pages 833–841.

Zhang, Q. and Chan, A. B. (2019). Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In *CVPR*, page 8297–8306.

Zhang, Y., Zhou, C., Chang, F., and Kot, A. C. (2019). A scale adaptive network for crowd counting. *Neurocomputing*, 362:139–146.

Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597.