Active Learning for High-Density Crowd Count Regression

Jennifer Vandoni, Emanuel Aldea and Sylvie Le Hégarat-Mascle
SATIE - CNRS UMR 8029
Paris-Sud University, Paris-Saclay University, France
{jennifer.vandoni, emanuel.aldea, sylvie.le-hegarat}@u-psud.fr

Abstract

Efficient crowd counting is an essential task in crowd monitoring, and significant advances have been made in this field recently by counting-by-regression techniques. We propose in this work a learning-to-count strategy with a generic detection algorithm which benefits from a counting regressor in order to identify crowded subregions with inadequate head detection performance, and to improve their representativeness in the training set. A straightforward but crucial step is proposed in order to take into account perspective correction within the proposed framework. An evaluation on Makkah images with medium to very high densities demonstrates the effectiveness of our algorithm and its capability to reach a count error of less than 5% in this difficult setting.

1. Introduction

Pedestrian scene analysis in videos has become a very active topic of research in the last decade due to a number of concurring factors: advances in the underlying detection algorithms, the availability of more computational power and a higher demand to automate as much as possible tasks in public security and situational awareness. Among various subtopics, counting and especially crowd counting has received growing attention recently.

Counting by detection (CD) is a straightforward way of approaching counting, by delegating the task to a detection algorithm [11, 10]; in this case the binary output of the detection is integrated directly in order to provide the count. The main inconvenience of this approach is that relying on the detector to provide crisp detections requires to perform thresholding and non-maximal suppression, which are not adapted in the case of close or partially occluded objects. Counting by regression (CR) approaches aim to map image characteristics to the number of objects being present [15, 8, 5]. In occluded scenes, CR methods have been shown to be better suited, but their main limitation is that they do not infer the actual object locations (although their output may be used as a prior for guiding detection and tracking [20]).

A field which illustrates the interest of counting (even though object localization is not addressed, i.e. using CR) is the physical modeling of high-density crowds. Individual tracking provides the maximum of information about the spatio-temporal crowd system state, but a number of macroscopic models [3] require only a local density estimation, which is in all respects within the reach of CR.

Motivated by this application, we propose in this paper a generic learning strategy supporting CR in high-density crowd images, which addresses in original ways the main difficulties specific to this context: the limited size of objects, the frequency and the high degree of occlusions, the specificity of the data (high class imbalance, user annotations being performed in batch) and the sensitivity to the variation in size of the objects due to the geometry of the scene. The present work falls into a category of learn-to-count algorithms [15, 12, 5, 23], which refine counting iteratively using some feedback based on the current performance and possibly on user additional input. In order to tackle practical scenarios related to high-density crowd analysis, we contribute in the following aspects: **a)** for learning, we propose a criterion for selecting the most informative training samples from a pool of strongly unbalanced dot annotated data, **b)** we propose a mechanism to integrate a counting objective in retraining efficiently a high-density head detector, and **c)** we propose a solution to integrate a correction for perspective distortion into the regression.

2. Related work

Let us introduce the state-of-art for the two aspects that are involved in our work (counting by regression and active learning), stressing the limitations of the presented approaches for our case study, namely high-density scenes.

CR in high-density images Count estimation methods, and specifically CR, have developed significantly in the last decade, and a number of surveys are available [21, 16]. Initial work relying on CR and on variations based on region clustering [6], motion patterns [19] etc. was not aimed

at tackling high-density crowds. Count estimation in small crowds is performed in [18] relying on accurate camera calibration and area of projection. However, this strategy is ideally suited for crowds that may be divided into groups of relatively homogeneous densities. In [17] self-organizing neural maps are used to infer the crowd density from image texture, but the task is aimed at identifying the correct density range rather than accurate counting, particularly in a high-density crowd.

Under a regularized risk framework, the objective of CR is to recover a linear transform defined by a parameter w which maps an estimated density F to a user-specified ground truth G. This may be formulated as:

$$\hat{w} = \arg\min_{w} \left(w^2 + \lambda \sum_{i=1}^{N} \mathcal{D}(G(\cdot), F(\cdot|w)) \right), \quad (1)$$

where $\mathcal{D}$ is a distance measure and λ a scalar weighting parameter.

In [15], the authors address a major limitation of image-level regressors based on Eq. (1) when using as distance measure an absolute or squared difference between the sums over the entire images. Such simple approach requires a large variety of image samples during training. Therefore, the authors of [15] propose a new distance called MESA, which takes into account the mapping penalty for all the possible B within the corresponding 2D box space $\mathcal{B}$ of mapping and ground truth areas:

$$\mathcal{D}_{MESA}(G, F) = \max_{B \in \mathcal{B}} \left| \sum_{\mathbf{p} \in B} G(\mathbf{p}) - \sum_{\mathbf{p} \in B} F(\mathbf{p}) \right|. \quad (2)$$

The significant strengths of this distance are an improved robustness to additive local noise, as well as the ability to exploit not only the ground truth count but also its spatial layout. Further work adapted the initial idea in order to alleviate the computational cost required to solve the optimization program of [15], by relying on regression trees [12]. Although the MESA distance has been used extensively for CR, a number of limitations have been underlined, such as a bias towards tuning the hyperparameter w with respect to a specific density level [13]. Also, for complex scenes a linear regression model may be too simple to map the input data to the count function.

Learning to count Active or interactive learning have been used in relation with CR as follows. In [2], the user is requested, and assisted with two density visualization techniques, to identify bad estimations. However, for high-density images the visualization assistance is not adapted, and also due to the extreme head proximity in crowd images, the only acceptable way of annotation is by performing it progressively on entire views. In [7], the authors address the problem of identifying informative samples for annotation in unlabeled images, whereas we have labeled data with only a subset used for the learning step and our main problem is to increase the size of the learning dataset in a tractable manner by selecting new data that will improve the further estimations. Closer to our proposed approach, the authors of [9] adapt regression in order to back-propagate the importance of samples as weights in the regression algorithm. The underlying assumption is that smaller errors point to samples which capture consistently the model and which should be weighted higher. However, in this study, our problem is to define an efficient selection of the yet unused labeled samples in order to increase the CR performance.

Cross-scene counting Recent works based on deep convolutional architectures [25, 24] have reached impressive results in cross-scene counting and detection tasks. Beside the computational cost, the performance and the improved robustness to scene variations come at the price of requiring significantly more annotated data (two orders of magnitude more compared to our method). Active learning and the use of prior geometric information about the scene allow us to compensate for a smaller, easier to set up training set, and provide a lightweight solution for learning to count in a specific setting.

Our basic assumption is that in some contexts where MESA distance based CR is known to perform more deficiently, this behavior is not due to the regression step, but rather to the lack of some appropriated data to consider in the regression. Thus, we propose to mediate through a feedback loop the performance estimated during the regression step. This feedback aims to improve the quality of the input in areas where the image characteristics are unreliable.

With respect to existing image-level regressors, we consider that the MESA approach is better suited for high-density annotated images for the following two reasons. Firstly, as a L_∞ distance between combinatorial sub-area vectors of the ground truth and of the score map provided by the detector, the MESA distance is ideally suited for a feedback strategy which is aimed at identifying subareas where the input map should be improved. Secondly, many applications such as physical modeling of crowds rely on local density estimations, and through the set of boxes $\mathcal{B}$, the MESA distance considers all image scales in order to achieve better robustness of density estimation across the whole scale space.

3. Details of our method

In light of the limitations mentioned above, we propose to apply the MESA distance to the probabilistic output of a general detection algorithm, and use the subregion (box) with the most violated constraint provided by the regression in order to select new informative training examples for the detector. In this way, the potential nonlinearity between the feature space and the mapping is dealt with by the learn-

ing step, and the regression is used secondarily to pinpoint badly mapped image parts which can provide new valuable training samples. In this sense, the algorithm may be seen as an *objective-driven* active learning with the aim of count regression. Indeed, the objective itself (count regression in our case) is directly involved into the choice of the new training samples that will improve the estimations. This strategy addresses at the same time a fundamental problem faced by discriminative learning on high-density crowd images, where pixel-based training sample sets are large, but ambiguous and highly unbalanced. Human annotations are performed by clicking on the object center, but as the object size reaches only a few pixels and as the occlusions may cover more than half of the object, the pixel assignments become highly subjective and unreliable.

We consider a generic binary classifier which provides for each tested instance (pixel) $\mathbf{p}$ a score $s(\mathbf{p})$ representing the probability of $\mathbf{p}$ belonging to the positive class $P(y = 1|\mathbf{p})$. Our aim is to recover the scalar factor w which maps a density $F(\mathbf{p}) = ws(\mathbf{p})$ based on Eqs. (1) and (2).

Computing the MESA distance may be cast efficiently as a max 2D subarray problem, while determining w requires solving a convex QP with a combinatorial number of linear constraints in a tractable manner [15]. Concurrently with solving for the optimal w, we identify the most violated box $\tilde{B}$ corresponding to the maximal mapping error. This allows us to select inside $\tilde{B}$, using a criterion specific to the learning algorithm being used, the most informative samples that would improve at the next learning iteration the score in the critical area $\tilde{B}$. For illustrating our method, we rely on an SVM classifier, and the following paragraphs will detail the preparation of the training set and the selection strategy we propose.

Building a training set As in [15], we perform a dotted

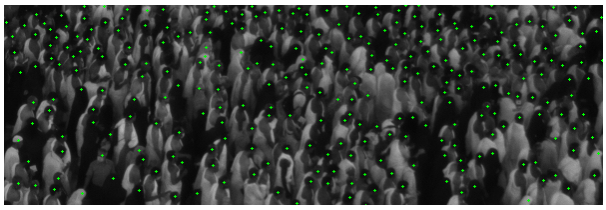

Figure 1: Patch with ground-truth dotted annotation

annotation in the head centers for the training images (Fig. 1). Then, we define the positive sample areas $\mathbf{P}_{all}$ a one pixel dilation of the annotations, in order to enlarge slightly the human clicks while avoiding to label as positive samples pixels which are on the border of a head and which can be misleading for the classifier. For the same reason, we define an exclusion zone up to 6px around the positive samples, and all the pixels present outside the exclusion zone are assigned to the negative label set $\mathbf{N}_{all}$. In this way, we encourage a detector with a peaked response, while at the same time selecting negative samples which do not confuse the classifier.

Active learning We adopt an uncertainty sampling approach, which iteratively requests the labels for the instances whose classes are the most uncertain, i.e. in the context of SVM, the instances which are the closest to the separation hyperplane [22]. Sample clustering may help in mapping the feature space more efficiently, at some computational cost. Since our potential training set is quite large, we adapt [4] which considers the *diversity* between samples. In particular, the authors propose a selection strategy which aims to reach a trade-off between (i) the minimum distance from the hyperplane and (ii) the maximum angle between the hyperplanes defined by each sample. Denoting I^* the pool of indexes of available samples with a distance from the hyperplane less than one, the training batch S is built by incrementally adding a new example x_t such that:

$$t = \underset{i \in I^* \setminus S}{\arg\min} \left(\beta \|f(x_i)\| + (1-\beta) \max_{j \in S} k^*(x_i, x_j) \right) \quad (3)$$

where $\|f(x_i)\|$ is the distance of the sample x_i to the separation hyperplane, and where, given the two sample hyperplanes h_i and h_j and the kernel function k, we have:

$$k^*(x_i, x_j) = |\cos(\angle(h_i, h_j))| = \frac{|k(x_i, x_j)|}{\sqrt{k(x_i, x_i) k(x_j, x_j)}} \quad (4)$$

The β parameter can be tuned to control the trade-off between the classical strategy which takes into account only the distance from the hyperplane and the new approach that combines it with the diversity measure.

Since it is prohibitively costly to compute angles among all the available instances in $\mathbf{P}_{all}$ and $\mathbf{N}_{all}$, we propose a greedy preliminary selection of a potential sample set. Denoting by H the learning batch size, we select the KH examples closest to the hyperplane by using a priority queue over the potential training set with a negligible computational overhead. Then we apply the exhaustive diversity search in terms of cosine similarity among these KH samples, by caching only a K^2H^2 element Gramm matrix. For our needs, we found that $K = 10$ is adequate, but higher values will promote more diversity with an increased computational cost.

Perspective correction Correcting for perspective distortion has been addressed in counting tasks [14], although it is not systematically implemented [2] since some additional information is needed about the camera intrinsic parameters and about the camera-to-ground relative pose. However, a detector which has been trained with examples of varying size provides similar pixel-level scores for identical objects which have different sizes in pixels due to the perspective change. This would affect significantly the MESA

Input: $Train\ set\ \mathbf{I}_{tr}, Pos.\ set\ \mathbf{P}_{all}, Neg.\ set\ \mathbf{N}_{all}$
$Regression\ and\ validation\ sets\ \mathbf{I}_r, \mathbf{I}_v$
$Init\ training\ set\ size\ M, batch\ size\ H$
Output: $detector \mathcal{L}, mapping\ w, count\ error\ \epsilon_{count}$
$\mathbf{P}, \mathbf{N} \leftarrow RandomSelect(\mathbf{P}_{all}, \mathbf{N}_{all}, M)$
repeat
$\mathcal{L} \leftarrow train(\mathbf{P}, \mathbf{N})$
$\tilde{B} \leftarrow MESA(Gtruth(\mathbf{I}_{tr}), \mathcal{L}(\mathbf{I}_{tr}))$
$\hat{\mathbf{P}}, \hat{\mathbf{N}} \leftarrow DiversitySamp(\tilde{B}, \mathcal{L}(\mathbf{I}_{tr}), \mathbf{P}_{all}, \mathbf{N}_{all}, H)$
$\mathbf{P} \leftarrow \mathbf{P} \cup \hat{\mathbf{P}}, \mathbf{N} \leftarrow \mathbf{N} \cup \hat{\mathbf{N}}$
$w \leftarrow MESA(Gtruth(\mathbf{I}_{reg}), \mathcal{L}(\mathbf{I}_{reg}))$
$\epsilon_{count} \leftarrow error(Gtruth(\mathbf{I}_v), w\mathcal{L}(\mathbf{I}_v))$
until $STOP$

Figure 2: Outline of the proposed approach.

hyperparameter w which could only settle for an inadequate compromise among the various sizes. Similarly to [12], we compute a perspective map $\mathcal{M}$ based on an accurate camera-to-ground pose estimation [1]. Then we are able to compensate the distortion by multiplying the detector score with the corresponding factor provided by the distortion map: $\hat{s}(\mathbf{p}) = \mathcal{M}(\mathbf{p})s(\mathbf{p})$.

Finally, Fig. 2 synthesizes the steps of our algorithm, while Fig. 3 provides a visual representation of the algorithm workflow where the count feedback is underlined in the training step.

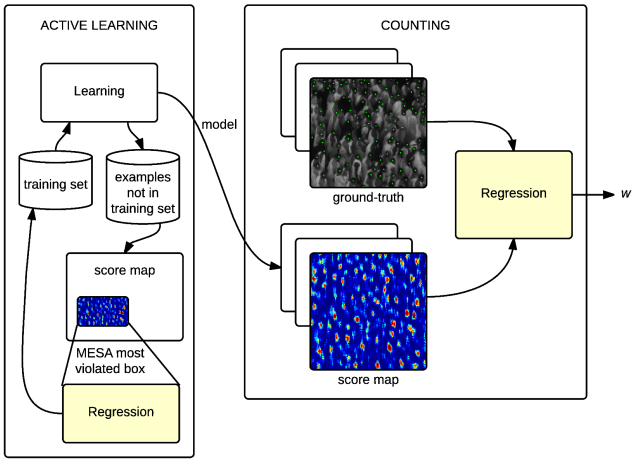

Figure 3: A visual representation of the algorithm workflow.

4. Experimental results

We validate the proposed active learning strategy using a linear SVM on a HOG descriptor with 1620 features. The HOG window size is set to 24px, which allows us to gather relevant information from the actual head but also from its close surroundings.

positive samples	negative samples	excluded samples
4055 px	662285 px	94076 px
0.53%	87.09 %	12.38 %

Table 1: Distribution of pixel samples in annotated images

Our dataset consists in images of high density crowds with an average of 800 heads. For a typical image, Table 1 shows the percentage of positive and negative examples, highlighting the data imbalance and the importance of finding an effective strategy to select significant samples in order to improve the score map for the counting objective.

We compared our new active learning approach with two widely used methodologies: the classical strategy which selects the closest examples to the separation hyperplane, from now on called *distance*, and the *diversity* strategy proposed by [4] explained above. In order to prove the effectiveness of active learning, we compared it also with a *random* strategy, which iteratively selects random, balanced examples from the pool.

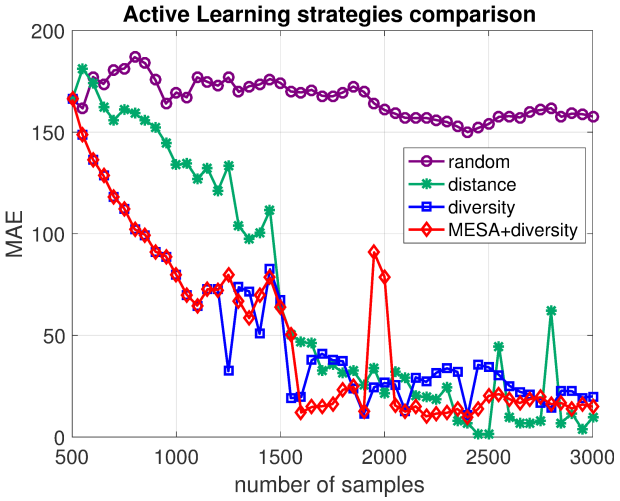

Figure 4: Comparison between different active learning strategies. The error drops immediately with the proposed MESA+diversity approach, and it remains stable towards the end.

As our training set is very unbalanced, metrics such as AU-ROC are unreliable for quantifying the learning performance. However, since counting is our main objective, we can directly perform counting on a validation set and use the final count error as a mean to assess the learning progress at each iteration. Figure 4 plots the Mean Absolute Error (MAE) for counting, with respect to the number of samples on the validation set $\mathbf{I}_v$, with perspective correction applied. The *random* strategy does not provide meaningful improvements as the training set becomes larger. On the contrary, the errors of all the active learning techniques

significantly drop from the beginning. In particular the *distance* approach improves slower, and presents some oscillations even towards the final iterations, while errors for the *diversity* strategy, and for the proposed approach called *MESA+diversity* drop immediately and then remain stable towards the end, highlighting the importance of the variety between the selected samples. It is possible to notice that for the first iterations the samples selected by the two methods based on *diversity* are the same. This happens because the box selected using the MESA distance as the most violated one is very large. Moreover, both learning and regression benefit from each other, and we highlight that MAE in the context of the counting task is a better performance metric, with respect to learning statistics which lose their applicability in presence of high data imbalance.

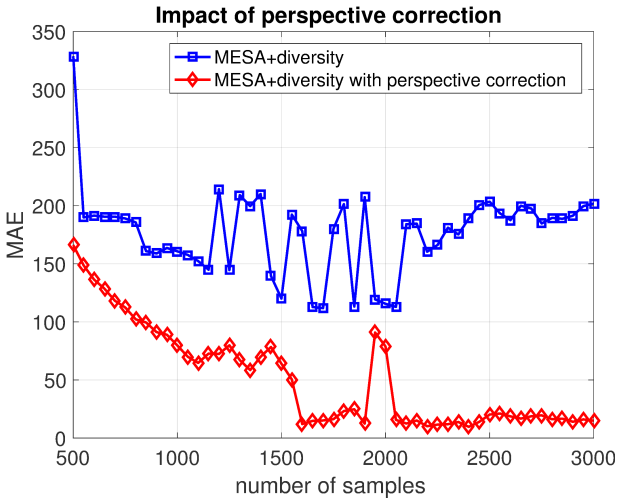

Figure 5: Impact of perspective correction on count estimation.

Figure 5 shows the importance of the perspective correction for the MESA regression, which compensates the head size variation with respect to the camera. The perspective correction step is crucial in order to obtain a low MAE and a stable behavior.

5. Conclusion and perspectives

In this work, we have shown how to exploit a highly unbalanced labeled set of head annotations in order to refine a crowd counting algorithm. Our approach is applicable to relatively small training sets made up of a few thousands of compact head annotations. Prior information about the geometry of the scene may be easily integrated as well into the algorithm through a perspective correction map. Overall, the proposed strategy is fairly easy to deploy for a given scene, and the results we get on images acquired at Makkah at peak times are very encouraging, with a count error of less than 5% on images which are difficult to annotate by human subjects. The perspective of our work are related to the use of more complex features and/or detection algorithms, in order to benefit fully from the nonlinearity of the detector score versus the mapping effectiveness of the MESA regressor. We also intend to investigate possible solutions for cross-scene counting based on our iterative learning strategy.

Acknowledgments

This work was partly funded by ANR grant ANR-15-CE39-0005 and by QNRF grant NPRP-09-768-1-114.

References

[1] E. Aldea and K. H. Kiyani. Hybrid focal stereo networks for pattern analysis in homogeneous scenes. In *ACCV 2014 Workshops*, pages 695–710, 2014.

[2] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman. Interactive object counting. In *European Conference on Computer Vision*, pages 504–518. Springer, 2014.

[3] A. Aw and M. Rascle. Resurrection of" second order" models of traffic flow. *SIAM journal on applied mathematics*, 60(3):916–938, 2000.

[4] K. Brinker. Incorporating diversity in active learning with support vector machines. In *ICML*, pages 59–66, 2003.

[5] J. Cavazza and V. Murino. People counting by Huber loss regression. In *ICCV Workshops*, 2015.

[6] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008.

[7] C. Change Loy, S. Gong, and T. Xiang. From semi-supervised to transfer counting of crowds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2256–2263, 2013.

[8] K. Chen, S. Gong, T. Xiang, and C. Change Loy. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2467–2474, 2013.

[9] K. Chen and J.-K. Kämäräinen. Learning to count with back-propagated information. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4672–4677. IEEE, 2014.

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2012.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[12] L. Fiaschi, U. Koethe, R. Nair, and F. A. Hamprecht. Learning to count with regression forest and structured labels. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2685–2688. IEEE, 2012.

[13] H. Idrees, I. Saleemi, C. Seibert, and M. Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2554, 2013.

[14] D. Kong, D. Gray, and H. Tao. A viewpoint invariant approach for crowd counting. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 1187–1190. IEEE.

[15] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*, pages 1324–1332, 2010.

[16] C. C. Loy, K. Chen, S. Gong, and T. Xiang. Crowd counting and profiling: Methodology and evaluation. In *Modeling, Simulation and Visual Analysis of Crowds*, pages 347–382. Springer, 2013.

[17] A. N. Marana, S. A. Velastin, L. d. F. Costa, and R. Lotufo. Automatic estimation of crowd density using texture. *Safety Science*, 28(3):165–175, 1998.

[18] P. Morerio, L. Marcenaro, and C. S. Regazzoni. People count estimation in small crowds. In *Advanced video and signal-based surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 476–480. IEEE, 2012.

[19] V. Rabaud and S. Belongie. Counting crowded moving objects. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 705–711. IEEE, 2006.

[20] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert. Density-aware person detection and tracking in crowds. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2423–2430. IEEE, 2011.

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

[22] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *ICML*, pages 839–846. Citeseer, 2000.

[23] S. Seguí, O. Pujol, and J. Vitria. Learning to count with deep object features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 90–96, 2015.

[24] R. Stewart, M. Andriluka, and A. Y. Ng. End-to-end people detection in crowded scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[25] C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 833–841, 2015.