IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 09, Issue 6 (June. 2019), ||S (IV) || PP 01-07

Kernel Based Object Counting Using Density Estimation

Parvathy Mohan G.[1], Riya John[2]
Department of ECE, MBCET, Thiruvananthapuram, India
Corresponding Author: Parvathy Mohan G.

Abstract: Object counting focuses on the prediction of the number of objects in a given image of interest or a frame from a video. This can be further used for obtaining the object statistics in a given counting scenario. There are many approaches in count estimations. Here density estimation based approach is considered. But the performance of density estimation based method seems to decrease when new counting scenarios are considered which can be solved by manifold based approach. In this paper a patch based method which makes use of kernels is presented. Here importance is given to techniques that make use of less number of training samples. In this paper various object counting scenarios are discussed and apart from the kernels used for counting purpose, exponential kernel is explored in this paper.

Keywords: Density estimation, density-map reconstruction, kernel-based, kernel function, object counting

Date of Submission: 11-06-2019 Date of acceptance: 27-06-2019

I. INTRODUCTION

The counting scheme deals with evaluating the count of objects of interest from an image or from video frame. It has relevance in many real-life applications including cell counting for bio-medical applications, crowd monitoring in surveillance systems, and taking wildlife survey or estimating the tree count from an aerial . Visual object counting is a standout amongst the most dynamic fields of research which concerns the areas of computer vision and signal processing that deals with estimating the object count in an image or video frame. This can also be used for obtaining the object distribution information in a given image or scene. This procedure can thus be utilized for various counting scenarios, e.g. cell counting in the field of biomedical imaging, for taking wildlife survey or census, and crowd monitoring and counting in public areas. At present object counting is a demanding task of correctly obtaining the count of objects for dense or crowded scenes .In the case of a crowded scene, the occlusions among objects will be abundant and results in miscalculations in counting. Also perspective distortions are most likely to vary significantly in different areas. Apart from this, there will be diverse object distribution information. All these issues have affected the execution of currently used object counting strategies. There has been an approach (Wang et al., 2018) based on manifold assumption for visual object counting is performed. In this method the manifold assumption is taken such that if image patches are found to be similar in nature then such image patches shares similar object densities too. As initial step, for a given image patch, it’s local geometry is expressed linearly using its neighbors by making using of a training set of image patches. Then the object density of this particular image patch can be regenerated after applying locally linear embedding while conserving its local geometry.

In (Lempitsky et al., 2010) a supervised learning methodology is implemented which can be used for visual object count estimation schemes, such as cell counting in the field of biomedical imaging, pedestrian count estimation etc. Initially the training images are user annotated using dots .In this method, the main aim is to correct estimation of the object count. Here a loss function is used which can be evaluated effectively using a maximum sub-array algorithm. This loss function is adequate for learning process. Here learning is then exhibited as a convex quadratic program which is solvable using cutting-plane optimization technique. This methodology is incredibly adaptable because it can take any features which are visual in nature that is specific to a particular domain. After training for just one time, this technique gives exact count of objects .Also this method only needs small execution time for extracting features. Hence it’s an apt choice for applications that demands real-time processing speeds or for handling with extensive volume of visual data. In (Antonini, 2006) clustering techniques are utilized for automatic estimation of pedestrians in video frames is implemented. The output of detection systems that estimate the count of targets is given as input to the system. Clustering techniques are carried out to the resultant trajectories so as to scale down the bias between the number of trajectories and the original count of targets. Diverse data depictions and different distance measurements are correlated, using a general hierarchical clustering scheme. A binary search tree which is multidimensional in nature is developed in this method (Bentley, 2006). This is typically called k-d tree in which the dimension is k.

This k-d tree is basically a data structure which is used for obtaining data according to corresponding searches. The search tree described here is found be efficient for various storage needs. The benefit of using this tree structure is that various kinds of searches can be handled by using a single tree structure. Several kinds of utility algorithms are also specified in this method. This algorithm is superior to currently existing algorithms for performing the same task. A theoretical approach is presented in this method for handling a general intersection query. For visual object counting, as part of hierarchical clustering this k-d tree is used as binary search tree. In (Huang, 2018) a density based method which make use of body part map and structured density map is described.

II. OBJECT COUNTING STRATEGIES

Typically, most of the existing schemes for visual objet counting can be grouped into three sets. They are detection based, global regression and density estimation based method.

a. Detection Based Method

Object detection is framed most commonly as a binary sliding window scheme for classification. In this technique the first step is sliding a window having fixed size over an image. Then based on some non-maximum suppression procedure, bounding boxes are localized around objects. For pedestrian counting, detection of humans from images is done by considering shape as a dominant feature(Lin et al., 2010). This is because there will be considerable variability in appearance. For videos, motion is used as a favorable feature for detecting humans. Two types of object detectors are there .One is generative and the latter one is discriminative approach. For efficient shape matching, in the case of generative approach a tree-based data structure is constructed. After template matching and nearest neighbor search, a vote is evaluated for each detection window. Adaboost classifiers and support vector machines are most commonly used discriminative approaches.

To match individual objects in images various kinds of detectors are used. (Li et al., 2008)Histogram Oriented Gradients based detectors are used to detect objects within foreground areas. In the case of pedestrian counting, part detectors are used for detection of local human body parts. Finally the result is taken by combining the outputs from part detectors in order to estimate people detections. This method when used in crowd counting gives best results if the objects in the scene are well separated and free from occlusions. If the objects are closer or subjected to occlusion then this methods fails since there will be miscalculations in count estimation.

b. Global Regression Method

Object counting using global regression methods basically learns the mapping between local features in the images and count of objects of interest. This method works better for crowded scenes compared to the detection based counting method. Different types of low level features like, textures, edge information, and segment shapes are used in this method. Apart from that,(Antoni et al., 2012) various regression algorithms like Gaussian process regression, linear regression, bayesian regression, and ridge regression are used frequently. This methods make use of the information regarding object counts, while, the object distribution details are ignored. The count estimation by global regression method can provide better estimation than detection based method by using quick training and by subsequent testing. But this method heavily depends on feature engineering. Hence this method cannot give accurate spatial information.

Regression-based methods basically deal with regressing from global image features to the whole image or in some cases, input patch count. All location information is discarded. Also mapping of local features to object blob count is done according to the output after segmentation. Regression based method ignores the information regarding location of the objects within a particular area of interest in an image. So this method cannot be used for applications needing object location. Extraction of features that are apt for a particular application is a crucial part of regression-based methods.

c. Counting By Density Estimation

Density Map Estimation deals with estimation of density at each pixel location of an image. This method provides spatial information about object location. From the view point of pedestrian counting, this method has applications in the field of safety and surveillance. Because very great density of crowd at a particular region in an area can be dangerous .In this method ,for modeling the spatial information of objects in images, the latent density distributions of objects is reformulated as an intermediate ground truth which is density estimation. In this method, from low level features, the density values are estimated. This method thus shares the benefits of global regression-based methods, while it also preserves spatial statistics.

As first step, (Lempitsky et al., 2010)using two dimensional Gaussian kernel , the density map is generated according to user annotated points. Then a linear regression is learned between an image and its corresponding density map. The density estimation method is usually comprised of three typical steps: Firstly using user annotated training images, the ground truth density map is generated. Then from the images the local features are extracted. As final step, for learning the mapping between the local features and their corresponding density map a linear regression model is applied. After that, the regression model which is learned can be applied for the estimation of density map of any image. The object count is evaluated by performing integration over the density map. Deep neural networks and random decision forest techniques are most commonly used approaches in density estimation.

III. METHODOLOGY

For conventional density estimation based object counting approaches, given an image X, as the first step the corresponding density map X_d is evaluated. Then the object count can be evaluated by computing the integral over X_d. Based on a new context, in this method, a unique methodology is taken for estimating the density map X_d. Using manifold assumption and learning techniques, this method estimates object density.

When spatial space is taken into consideration, images and their corresponding density maps shares exact object location information. Also it is known that that many number image patches share similarity in the counting scenario. This indicates that periodic patterns are likely to occur everywhere in the case of natural crowd scenarios such as pedestrians and birds. Based on these two observations, the manifold assumption is taken in the counting scenario as follows: the objects in the images and their density maps share same features. Hence, the image patches and density maps which shares similar local geometry are considered to be belonging in two manifolds.

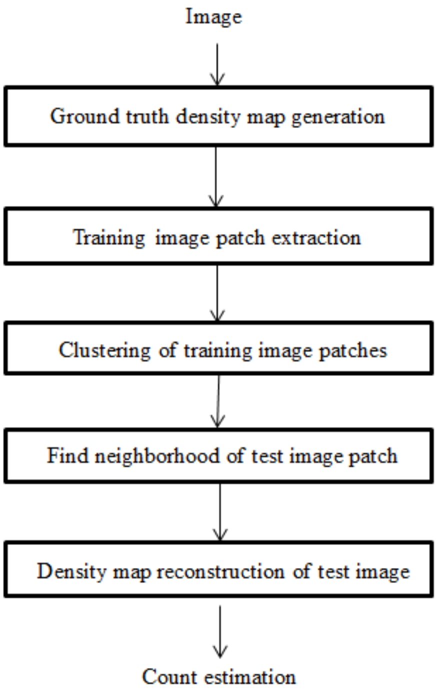
Image
Ground truth density map generation
Training image patch extraction
Clustering of training image patches
Find neighborhood of test image patch
Density map reconstruction of test image
Count estimation

Figure 1: Block diagram of count estimation

As shown in Fig 1 , the first step is to compute the ground truth density map as in (Lempitsky et al., 2010). The user annotated location of objects are basically discrete two dimensional points denoted in the image. As a task of varying the object locations in a continuous manner, the location map of objects for the images is kernelized. Hence the object distribution function will be smoothened. Now consider a group of N user annotated images which are denoted as $I_1, I_2, \ldots, I_N$ .These images are already allocated with object locations. As a result, the ground truth density maps are generally expressed as a sum of 2D Gaussian kernel functions of the object locations , as:

$$I_d^i(z) = \sum_{U \in U^i} \mathcal{N}(z; U, \sigma^2 1_{2x2})(1)$$

where the ground truth density map of an image I is denoted as I_d, pixel index is denoted by z, i is the image index, user annotated dot is denoted by U, and U^i is a two dimensional points set denoting all location of objects in the image I_i. Besides, the 2D Gaussian kernel used is normalized with variance σ^2 .Variance of N is utilized for smoothing the local distribution, and its value is set according to the object size. Generally it is roughly half the size of objects under consideration. In fig-2 an image from cell dataset and its kernelized ground truth density map is shown.

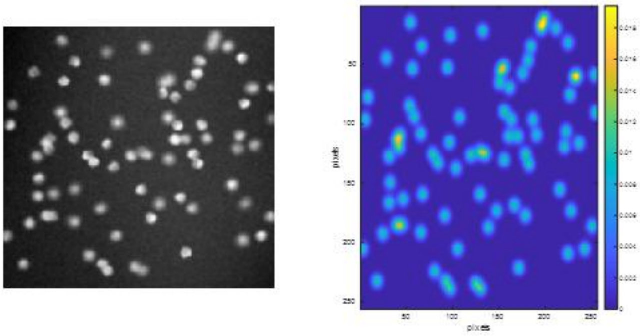

Figure2: An image from cell dataset and corresponding ground truth density map

Next step is training image patch extraction. After that extracted training patches are clustered using k-means hierarchical clustering as in (Wang et al., 2018).Then neighborhood of the test image patch is found using minimum centroid distance of clusters. After the density-map reconstruction, the count is estimated by taking summation of density values.

Consider x ,an image patch which is extracted from image X, and let $x_d$be its corresponding density patch. The user annotated training images can be denoted as I^i where i = 1,2,...,N. Let the image patches extracted be denoted as Y = $\{y^1,y^2,...,y^M\}$, in which $y^i \in \mathbb{R}^{q1}$. Hence for the given image patches, their corresponding set of the density patches can be given as Y_d where each element corresponds to density patches extracted from training images. The objective with this method is the estimation of x_d for a particular image patch x.

According to the assumption in (Wang et al., 2018) that image patches and their respective density patches belongs to the same manifold, x and x_d shares similar local geometry. That leads to the fact that if x can be described using its neighbors in a peculiar way for capturing the locality in images, then x_d can also be expressed using its neighbors in an equivalent way. The correspondence between them which is achieved on the local geometry between x and x_d can be denoted as:

$$X = Dw \qquad (2)$$

$$x_d = D_d w \qquad (3)$$

where D denotes subset which comprises of T nearest neighbors of x which is taken from Y , D_d denotes subset of density patches that corresponds to D, and w denotes weight vector that characterizes the local geometry of both x and x_d. w can be jointly computed from (x, D) and (x_d, D_d) theoretically. In practical sense, x_d is not known to us and hence it should to be predicted from x. Aim is to decrease the linear reconstruction error for determining w. The analytical solution for this optimization problem as in (Wang et al., 2018) is expressed as

$$w^* = \frac{1}{z}(D^T D + \lambda I)^{-1} D^T x \qquad (4)$$

As described in (Wang et al., 2018) a non-linear mapping is applied for mapping x to a higher dimensional space using mapping function φ. As described in (Alpaydin, 2014), w is computed by making use of kernels in new high dimensional space.

$$w^* = \frac{1}{z}(G + \lambda I)^{-1} k(D,x) (5)$$

Here G is the gram matrix and k(D, x) is the kernel relating D and x. Here G corresponds to the gram matrix and k(D, x) corresponds to the kernel which relates D and x. The reconstruction of the patch density map is approximated as

$$x_d \cong Ek(D,x) \quad (6)$$

Where E is the Embedding matrix. After clustering of the training patches and assigning centroids to the training patches, the cluster index i that corresponds to the centroid which is having minimum distance with a given test patch is calculated. Then the embedding matrix of thisith index is given by :

$$E^i = C_d^i (G + \lambda I)^{-1} \qquad (7)$$

C_d^i corresponds to the density patches of elements in ith cluster. Thus there are basically two steps in this method. In the first step,for the extracted test patch the neighborhood of ith cluster is found. Then the reconstruction of density patch corresponding to the test patch can be done by making use of the embedding matrix corresponding to the ith cluster. Here i is the cluster index whose centroid is having minimum distance with the test patch. Apart from the kernels used in (Wang et al., 2018) exponential kernel is applied in this paper.

II. KERNEL FUNCTION

When data points are mapped to a high dimensional space, representing data in this space is computationally difficult. Hence instead of representing the data points, a similarity measure is computed in the high dimensional feature space. This is known as kernel trick. The high dimensional space in which mapping is carried out could be of infinite dimension and it will be not feasible to compute. Then linear algorithms are applied that make use of this similarity measure. Dot product is used as similarity measure. Thus kernels basically act as a bridge from linearity to non-linearity.

Given a set of data points there are methods for finding a linear relation between them. But if the data points are non- linear in original dimensional space, this will not be possible. A kernel function is used to map a certain data point under consideration to a high dimensional space. This is established on the assumption that data points not separable in a given dimension would be well separated in a higher dimensional space. In the high dimensional space, linear relations exists among data ,hence a linear algorithm can be applied for this data in high dimensional space.From the given original dimensional space, a kernel function takes input vectors and gives dot product of the vectors in required high dimensional feature space .Given two vector x and z in a given dimensional space X i.e. x , z ϵ X and Φ is the mapping function where Φ : X →R^N, then kernel function is given by

$$K\,(x\,,z) = \langle \Phi(x).\Phi(z) \rangle \qquad (8)$$

Exponential kernel is given by:

$$K(x,y) = exp(-\frac{\|x-y\|}{2\sigma^2}) \qquad (9)$$

Here σ is set according to the problem under consideration.In this paper σ is set to 2.2.

IV. PERFORMANCE EVALUATION

In this paper as in (Wang et al., 2018) Mean Absolute Error(MAE) is used as performance evaluation measure. It is given by:

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left|\mathrm{C_i} - \widehat{\mathrm{C_i}}\right| \qquad (10)$$

Here $\widehat{\mathrm{C_i}}$denotes the true count and $\mathrm{C_i}$ is the count which is estimated and N is the number of images used for testing.

V. RESULTS AND DISCUSSION

a. Experimental Setup

In this paper four datasets are used. They are listed in Table 1. In cell dataset from the first 100 images, 32 random images are taken for the purpose of training. Remaining 100 images are taken for testing. For honeybee dataset, 16 images from first 68 images are used for training; while remaining are used for testing.16 images from the first 69 images from the fish dataset is taken for training. Rest of the images is taken for testing. For seagull dataset, first image is utilized for training and the second image is used for testing. For all datasets, 5-cross validation is carried out for finding MAE. Here patches of size 6x6 with a step size of 3 are used. Also the cluster size is taken as 256.For the experimental purpose blue channel of the images is taken.All the experiments are carried out in MATLAB R2018b.

Table 1: Dataset of images

Dataset	Resolution	No: of images
Cell	256 x 256	200
Seagull	624 x 964	3
Honeybee	640 x 480	118
Fish	300 x 410	129

b. Results

In Table 2 the results using exponential kernel is compared with Kernel based Manifold Visual Object Counting(KM-VOC) (Wang,2018) using RBF(Radial Basis Function) and laplacian kernel.In density+MESA(Lempitsky et al., 2010) count is estimated by making use of MESA(Maximum Excess over SubArrays) distance. In density+MESA as well as in codebook+RR(Ridge Regression)(Arteta,2014) ,real valued density function of pixels in the image is estimated by mapping local features of the image to its density map.In Figure 3, 4, 5and 6,a sample image and corresponding reconstructed density map of cell ,seagull , honeybee and fish datasets are shown respectively.

Table 2 : Results on datasets

Method	Cell	Seagull	Bee	Fish
	MAE	MAE	MAE	MAE
Density+MESA	3.5	10.4	3.8	2.6
Codebook+RR	3.5	12	4.2	4.0
KM-VOC(RBF)	3.3	6.8	2.8	6.8
KM-VOC(Laplacian)	4.0	3.2	3.7	2.9
VOC-Exponential	4.01	2.31	1.62	1.72

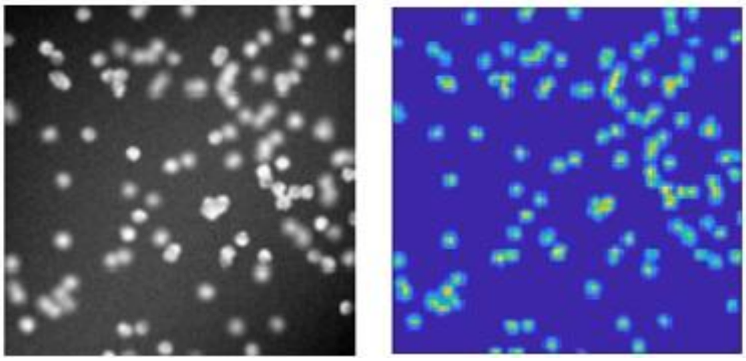

Figure 3: An image from cell dataset and corresponding reconstructed density map

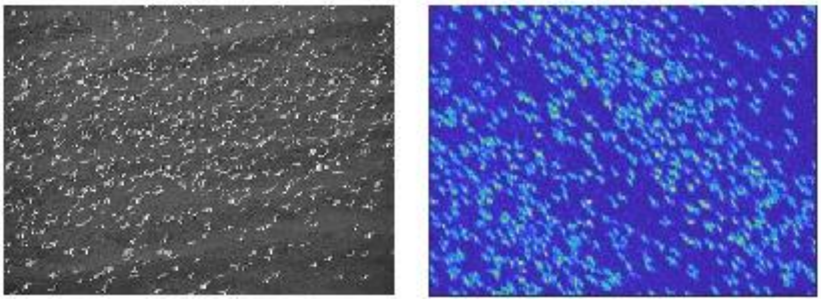

Figure 4: An image from seagull dataset and correspondingreconstructeddensity map

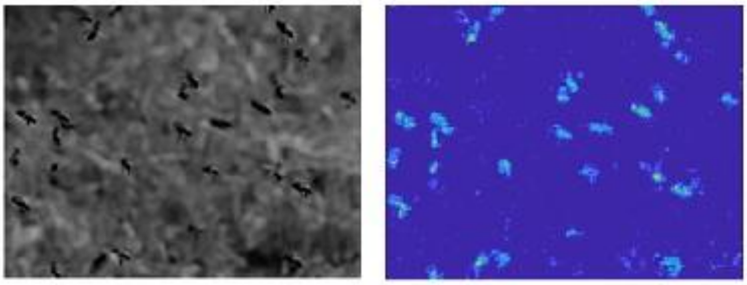

Figure 5: An image from honeybee dataset and correspondingreconstructeddensity map

From the results shown in Table 2, it is clear that object counting using exponential kernel gives better results compared to other counting methods, except for cell dataset. As given in Table 1, here only few images is used for training. This gives the significance of using raw image patch as feature in this method as in (Wang et al., 2018).

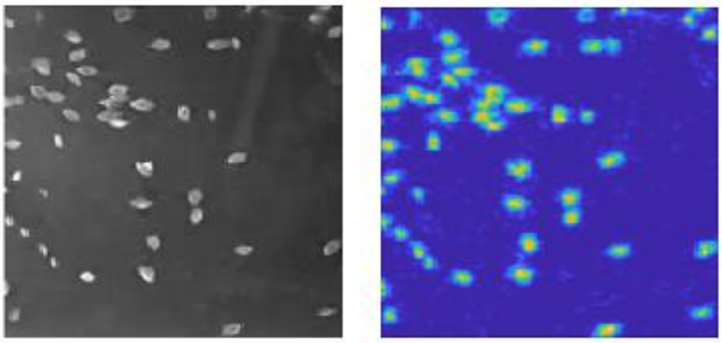

Figure 6: An image from fish dataset and correspondingreconstructed density map

VI. CONCLUSION

A visual object counting technique using density estimation is implemented. Also this method utilizes the scheme of local linear embedding for regenerating the density maps by utilizing a linear representation which is present in the neighborhood of image. In addition to this, to build adequate neighborhood and to solve the drawbacks in the local representation for taking count of objects in images with complex background, a mapping which is non-linear is employed along with kernels for this method. From the results obtained, it is clear that only few ranges of image samples are needed for training and the results attained were promising. In case of applications for which only relatively small datasets are available, then this method can be utilized. Because in such cases there will be restricted or small number of training data and hence CNN(Convolution Neural Network) based techniques which make use of large range of training dataset cannot be used. From the results obtained, it is clear than exponential kernel is giving better results for the given datasets than RBF and laplacian kernels.

REFERENCES

[1]. Yi Wang and YuexianZou and Wenwu Wang(2018), Manifold-based Visual Object Counting, *IEEE Trans.Image Processing*., 27(7), 3248 - 3263

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

[3]. G. Antonini and J.-P.Thiran(2006), Counting pedestrians in video sequences using trajectory clustering,*IEEE Trans. Circuits Syst. Video Technol*., 16(8),1008–1020

[4]. J. L. Bentley(1975), Multidimensional binary search trees used for associative searching,*Communications of the ACM*, 18(9),509–517

[5]. SiyuHuang,Xi Li (2018),Body Structure Aware Deep Crowd Counting, *IEEE Transactions on Image Processing*,27(3), 1049-1059

[6]. Z. Lin and L. S. Davis(2010), Shape-based human detection and segmentation via hierarchical part-template matching,*IEEE Trans. Pattern Anal. Mach. Intell*., 32(4), 604–618

[7]. M. Li, Z. Zhang, K. Huang, and T. Tan(2008), Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection, *Proc. IEEE ICPR*, 1–4.

[8]. Antoni B. Chan and NunoVasconcelos(2012),Counting People With Low-Level Features and Bayesian Regression, *IEEE Transactions on Image Processing*,12, 2160–2177

[9]. E. Alpaydin, Introduction to Machine Learning. MIT press, 2014.

[10]. C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman(2014), Interactive object counting, *European Conf. Comput. Vis. (ECCV)*,504–518.

IOSR Journal of Engineering (IOSRJEN) is UGC approved Journal with Sl. No. 3240, Journal no. 48995.

Parvathy Mohan G.. “Kernel Based Object Counting Using Density Estimation.” IOSR Journal of Engineering (IOSRJEN), vol. 09, no. 06, 2019, pp. 01-07.