EXAMPLE-BASED VISUAL OBJECT COUNTING FOR COMPLEX BACKGROUND WITH A LOCAL LOW-RANK CONSTRAINT

*X.L. Huang, Y.X. ZOU *, Y. Wang*

ADSPLAB/ELIP, School of ECE, Peking University, Shenzhen 518055, China
*Corresponding author: zouyx@pkusz.edu.cn

ABSTRACT

Visual object counting (VOC) is important in many real-world applications. Our previous work approximated sparsity-constrain example-based VOC (ASE-VOC) works well with insufficient training data. It assumes that image patches share the similar local geometry with counterpart density maps, and then the density map of the image patch can be estimated by preserving such geometry. However, ASE-VOC has a weak constraint for data structure and experiments reveal that the performance of ASE-VOC degrades when facing with complex background. To solve this problem, we proposed a novel local low-rank constrained example-based VOC (LLRE-VOC) method. Because local low-rank constraint can choose the samples belonging to the subspace that lies closest to the test samples. Even with complicated data structure, LLRE-VOC can guarantee the patches selected share similar structure with input patch. Extensive experiments conducted on public benchmarks demonstrate the superior performance of our proposed LLRE-VOC method.

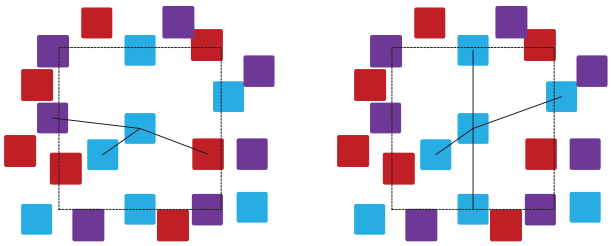
Index Terms— Visual object counting, complex background, local low-rank, density map estimation, example-based

1. INTRODUCTION

The task of Visual Object Counting (VOC) is to label an image with the exact object counts. In recent years, it has been widely applied in fields such as crowd analysis[2, 5-7], city resource management[8], public security[9] and wildlife census[10]. VOC has shown its great value in computer vision.

There are primarily two mainstream types of object counting techniques in a supervised way: one is based on global regression [2, 5-7, 11] and the other is density estimation [1, 3, 12, 13]. For global regression based method (GR-VOC), they learn an intrinsic mapping between image global features and their corresponding counts (in scalar form). These methods discard the location information of the objects. Moreover, the performance of GR-VOC depends on the well-design of feature heavily.

Compared with GR-VOC method, density map estimation based method (DE-VOC) takes full advantage of the spatial information and can provide object distribution information.



(a) ASE-VOC (b) our LLRE-VOC

Fig. 1. Conceptual Illustration of sample selection mechanisms. Patches filled with same colors are in the same subspace and have similar structure. (a) Sparsity constraint may select the patches in different subspace (b) LLRE constraint selects the patches in the closest subspace.

The main idea of the DE-VOC method is firstly proposed by Lempitsky who estimates a density function as a real function of pixels in an image regressed from dense local features of the image [1]. Based on Lempitsky's work, several algorithms [3, 12, 13] have been proposed to handle with different application scenes. Among those application scenes, how to deal with the insufficient training data is a tough problem.

In our previous work example based VOC (E-VOC) [3], we find that patches extracted from images share the similar local geometry with their corresponding patches extracted from counterpart generated density maps, therefore, by preserving such local geometry, the object density map can be reconstructed. E-VOC can work well with a few training images, but its result is unstable due to the effect of neighborhood size. To overcome this disadvantage, we introduce a sparsity constraint in our extended version which is called approximated sparsity-constrain example-based VOC (ASE-VOC). Experimental results show that ASE-VOC is able to give a good result when the background is clean or the foreground can be extracted. However, it is a challenge work to extract the foreground especially when the training data is insufficient (for example there is only one image). In this paper, we address the VOC problem for complex scenes with insufficient training data.

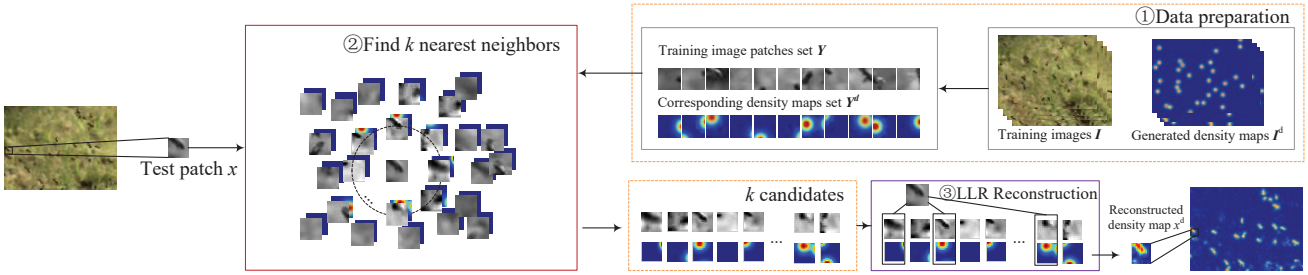


Fig. 2. Flowchart of proposed LLRE-VOC method. Dashed and solid boxes represents data and operations, respectively.

As discussed above, our previous proposed ASE-VOC [3] utilizes the sparsity constrain which does not consider the underlying structure of the data. As a result, when the training images are of complex scenes, the sparse constrain can't guarantee the sample selected are in similar structure, which degrades the performance. In this paper, we proposed a novel Local Low-Rank constrained example based VOC (LLRE-VOC). Instead of using sparse constraint, we exploit the property of local low-rank constraint for selection of samples. Work [14] shows that utilizing Locality-constrained Low Rank Coding (LLRC) in face recognition, the training samples used to reconstruct a given test sample can be chosen from just one class rather than a mixture of classes, thus enhances the classification accuracy. Motivated by it, we make an effort to apply local low-rank constrain to choose those samples that are in similar structure. Fig.1 shows the conceptual illustration of sample selection mechanisms in our proposed LLRE-VOC method and ASE-VOC method. Extensive experiments have been conducted to evaluate the performance of our proposed LLRE-VOC method on both simple and complex background datasets. Experimental results on public databases demonstrate the effectiveness of our proposed LLRE-VOC method. The flowchart of LLRE-VOC is shown in Fig.2.

The rest of paper is organized as follows. In next section, we will briefly introduce the general generation of density map. In section 3, we will give a description of the E-VOC problem formulation, followed by the presentation of our novel LLRE-VOC method. In Section 4, we will show the experimental results, and section 5 concludes our paper.

2. PRELIMINARIES

As our method is based on object density estimation, here we will introduce the generation of the ground truth density map briefly. Following the work [1], we assume that a set of N training images $\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_N$ are given. And for each training image $\boldsymbol{I}_i$, the objects interested are annotated with a set of 2D points $\mathbf{P}_i = \{ \mathbf{P}_1, \ldots, \mathbf{P}_{C(i)} \}$, where $C_{(i)}$ represents the number of objects which we are interested in image $\boldsymbol{I}_i$. Therefore, we define the ground truth density function to be a kernel density estimate based on the provided points:

$$F_i^0(p) = \sum_{P \in \mathbf{P}_i} N(p; P, \delta^2) \tag{1}$$

where P is a user-annotated dot and δ is the smoothness parameter. In our paper, δ we used here is set to be 6 in experiments.With the definition in Eqn. (1), the ground truth density map $\boldsymbol{I}_i^d$ of training image $\boldsymbol{I}_i$ is defined as

$$\forall p \in \boldsymbol{I}_i^d, \boldsymbol{I}_i^d(p) = F_i^0(p) \tag{2}$$

With the density map, the object count can be computed by integrating over the density map

$$c(\boldsymbol{I}_i) = \sum_{p \in \boldsymbol{I}_i^d} \boldsymbol{I}_i^d(p) \tag{3}$$

In our paper, the training image patches which are extracted from training images $\boldsymbol{I}_i, i \in \{1,2,\ldots,N\}$ are denoted as $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_M\}$. And the density maps $\boldsymbol{Y}^d = \{y_1^d, y_2^d, \ldots, y_M^d\}$ of corresponding image patches are derived from $\boldsymbol{I}_i^d, i \in \{1,2,\ldots,N\}$. For all training patches in $\boldsymbol{Y}$, the feature set can be represented as $\boldsymbol{Y}_f = \{y_{1f}, y_{2f}, \ldots, y_{Mf}\}$.

3. METHOD

3.1. Example-based VOC

In E-VOC, it supposes that the two manifolds formed by image patches and their density map patches, respectively, share similar local geometry. Such local geometry of a feature vector can be characterized by how the feature vector can be linearly reconstructed by its neighbors [15]. Given the feature of a test image patch x_f, the reconstruction weights of neighbors in feature space $\boldsymbol{Y}_f$ can be computed by minimizing the reconstruction error. Then we apply the reconstruction weights to the density maps of neighboring patches from $\boldsymbol{Y}^d$ and obtain the density map $\boldsymbol{x}^d$. This kind of method which uses the generalization of examples is named as example-base VOC (E-VOC) [3]. The formulation of E-VOC can be described as follows:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \|x_f - \boldsymbol{D}_Y \boldsymbol{w}\|_2^2 \tag{4}$$

$$\boldsymbol{x}^d \cong \tilde{Y}^d \boldsymbol{w}^* \tag{5}$$

where $\boldsymbol{D}_Y = [\tilde{\boldsymbol{y}}_{1f}, \tilde{\boldsymbol{y}}_{2f}, \ldots, \tilde{\boldsymbol{y}}_{kf}]$ is a training patch subset formed by the k nearest neighbors of x_f from $\boldsymbol{Y}_f$. $\tilde{\boldsymbol{Y}}^d = [\tilde{y}_1^d, \tilde{y}_2^d, \ldots, \tilde{y}_k^d]$ and $\tilde{y}_i^d$ is the density map of $\tilde{\boldsymbol{y}}_i$.

The Eqn.(4) computes the local geometry of x_f and then Eqn. (5) reconstructs the density map x^d by preserving the same local geometry. Due to the constrain with least square form, Eqn.(4) has an analytic solution and $\boldsymbol{w}$ can be calculated efficiently in Eqn.(6)

$$\boldsymbol{w} = (\boldsymbol{D}_Y^T \boldsymbol{D}_Y + \lambda \boldsymbol{I})^{-1} \boldsymbol{D}_Y^T x_f \tag{6}$$

Table 1. Mean absolute errors (MAE) for cell counting

Method	Feature	Validation	N=1	N=2	N=4	N=8	N=16	N=32
RR[2]	(1)	counting	67.3±25.2	37.7±14.0	16.7±3.1	8.8±1.5	6.4±0.7	5.9±0.5
KRR[4]	(1)	counting	60.4±16.5	38.7±17.0	18.6±5.0	10.4±2.5	6.0±0.8	5.2±0.3
detection[1]	(2)	counting	28.0±20.6	20.8±5.8	13.6±1.5	10.2±1.9	10.4±1.2	8.5±0.5
detection[1]	(2)	detection	20.8±3.8	20.1±5.3	15.7±2.0	15.0±4.1	11.8±3.1	12.0±0.8
Density learning[1]	(1)	MESA	9.5±6.1	6.3±1.2	4.9±0.6	4.9±0.7	**3.8±0.2**	**3.5±0.2**
E-VOC[3]	(3)	counting	20.5±11.8	**5.5±1.1**	**4.4±0.6**	5.2±0.6	5.0±0.2	4.8±0.5
ASE-VOC[3]	(3)	counting	8.1±3.6	5.9±0.9	4.9±1.1	4.8±0.7	3.9±0.3	3.6±0.1
LLRE-VOC	(3)	counting	**7.5±3.1**	5.8±0.8	4.8±0.7	**4.1±0.3**	3.9±0.1	3.7±0.2

(1) Dense SIFT+Bag of words; (2) Dense SIFT; (3) Raw data (extracted from blue channel)

Here the k nearest neighbors $\boldsymbol{D_Y}$ are searched by K-nearest-neighbors (KNN) algorithm.

3.2. Our proposed Local Low Rank Example-based VOC (LLRE-VOC) method

E-VOC works well with a small training set as it estimates density over generalization training patches. However, the result is unstable due to the fact that E-VOC fixes the neighbors size [3]. To overcome this disadvantage, we add the sparse constrain in extended version (ASE-VOC) to choose samples automatically. The formulation is as follows:

$$w^* = \arg\min_w \|x_f - \boldsymbol{D_Y} w\|_2^2 + \lambda \|w\|_1 \qquad (7)$$

Because ASE-VOC doesn't take structure of examples into account, thus, it cannot guarantee the selected examples are in similar structure especially when the data structure is complicated (for example the complicated background). As the Fig. 3 shows, the chosen samples in ASE-VOC method do not keep the similar data structure. To improve the performance of complex background, examples in similar data structure are favorable for reconstruction.

Recently, in face recognition, Locality-constrained Low Rank Coding (LLRC) [14] which chooses face images that belong to the same class that lies closest to the test face image by taking advantage of the low rank structure of data has achieved great success. Motivated by this, we introduce the local low-rank constrain into the E-VOC problem. And a novel local low-rank constrained example-based VOC (LLRE-VOC) method is proposed. The formulation of our proposed method is described as follows:

$$w^* = \arg\min_w \|x_f - \boldsymbol{D_Y} w\|_2^2 + \lambda_1 \|\boldsymbol{D_Y} diag(w)\|_* + \lambda_2 \|l \odot w\|_2^2 \quad (8)$$

where $w \in \mathbb{R}^k$ denotes the weights over these $\boldsymbol{k}$ vectors, the matrix $\boldsymbol{D_Y} diag(w)$ denotes the training sample used to reconstruct the input x_f and $l \in \mathbb{R}^k$ represents a vector which measures the exponential distance from x_f to each training sample d_i. Therefore, l_i is given by

$$l_i = \exp\left(\|x_f - d_i\| / \sigma\right) \qquad (9)$$

here, we normalize the value of l from 0 to 1. Parameter λ_1 and λ_2 are the regularization coefficients for trading off the structural similarity and locality.

Table 2. Mean absolute errors (MAE) for fly counting

Method	N=1	N=2	N=3	N=4	N=5	N=32
Density learning[1]	27.55	26.14	27.28	26.53	27.03	28.41
ASE-VOC[3]	16.50	12.31	11.71	11.04	10.69	10.06
LLRE-VOC	**11.84**	**9.39**	**9.31**	**8.97**	**8.59**	**8.58**

For solving the optimization problem, we convert the Eqn. (8) to the following formulation:

$$\min_w \|x_f - \boldsymbol{D_Y} w\|_2^2 + \lambda_1 \|Z\|_* + \lambda_2 \|l \odot w\|_2^2 \quad s.t.\ Z = \boldsymbol{D_Y} diag(w) \qquad (10)$$

Eqn. (10) can be solved by taking advantage of augmented lagrangian formulation. The detailed information about the solution of this model can refer to literatures [14, 16].

4. EXPERIMENTS

In order to demonstrate the effectiveness of our proposed LLRE-VOC method, we conduct experiments on three public dataset including Bacterial cell dataset [1], Fly dataset [17] and Honeybee dataset [17]. Fig.3 shows the example frames of three datasets. To compare with different methods, mean absolute error (MAE) is employed as the evaluation metric.

4.1. Bacterial cell dataset

The synthetic bacterial cell dataset [1] consists of 200 images with an average of 171 ± 64 cells per image. The resolution of each image is 256-by-256. Partial occlusion and image saturation exist in this dataset. Following the work [1], we only use the blue channel. Besides, the first 100 images used for training and the second 100 images for testing just as the same setting in [1]; the subset of N out of all training images is randomly selected. For each N (N =1,2,4, … ,32), the experiments have been repeated for five times and the mean absolute errors and standard deviations are calculated.

For our proposed LLRE-VOC method, the patch size is set to 4×4 both for training and testing, patch step is set to 2. The number of nearest neighbors k is set to 128.

The results are shown in Table 1. Compared with the classical detection based or GR-VOC methods, LLRE-VOC gives better results with no matter what the size of training samples is. Compared with the Lempitsky's method [1], our method offers more accurate estimation when training samples are insufficient, and provides competitive result when the size of training set grows. The mean absolute errors (MAE) produced by E-VOC (k is set to 5) method is unstable, which suddenly drops with N =4 and then rises with N =8 in Table 1, In comparison, our LLRE-VOC offers a more stable and accurate estimation. Compared with ASE-VOC, our method provides smaller mean absolute errors and standard deviations. Therefore, our proposed method has showed its superiority on benchmark dataset compared with existing mainstream methods.

4.2. Fly and honeybee datasets

Paper [17] provides four public datasets including fly, honeybee, fish seagull for detection. To further evaluate the performance of our proposed method on clean and complex background, we choose the fly and honeybee datasets which own the clean and complex background respectively. It is noted that here used features are just raw data extracted grayscale images and patch step is set to 4.

Fly dataset: contains 600 frames with an average of 86 ± 39 flies. The resolution of each frame is 648-by-72. Following the work [4], the first 32 images(1:6:187) are utilized for training and 50 images for testing(301:6:600).

For detailed performance comparison on insufficient training dataset, in this experiment, we use first N (N = 1,2,…,32) image for training respectively. Table 2 shows the results for fly dataset. From the table, we can find that both of our proposed method and ASE-VOC can achieve a satisfactory performance than density learning [1]. However, our LLRE-VOC method has a slight superiority over ASE-VOC, and shows a big improvement than ASE-VOC especially when there is only 1 training sample.

Honeybee dataset: contains 118 images with an average of 28 ± 6 honeybees per image. The resolution of each image is 640-by-480. First 32 images are used for training and last 50 images for testing.

To obtain the detailed performance comparison under complex background. We also use first N (N =1,5,…,32) images for training respectively. The results are given in Fig.5. It is obvious that ASE-VOC performs badly under the complex background. However, our proposed method still performs well and our MAE is almost the half of ASE-VOC's, which verifies the effectiveness of the proposed LLRE method with complex background. To validate the salient property of our method on patch selection, we visualize the patch selection result in Fig.3. The figure shows that our LLRE-VOC method choose those patches with similar structure. However, in ASE-VOC method, some dissimilar patches are chosen. As a conclusion, our proposed LLRE-VOC method is suitable for visual object counting with complex background.

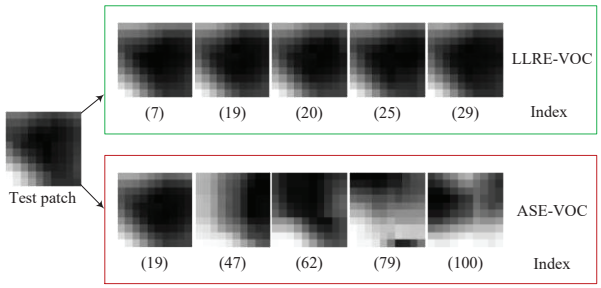
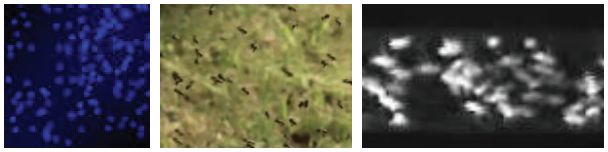


Fig. 3. The selection result for the input test patch. The numbers indicate the index of sample. The patches in red and green boxes are selected training samples.



(a) (b) (c)

Fig. 4. Examples of three public datasets. a) synthetic cells b) honeybees c) flies

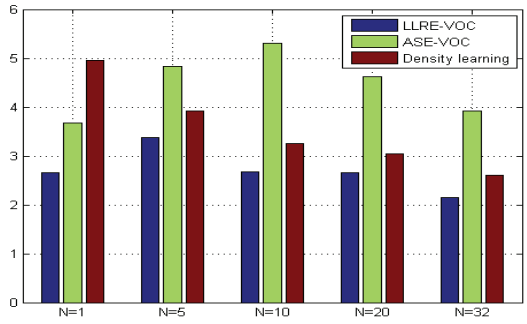


Fig. 5. Mean absolute errors (MAE) for honeybee counting

5. CONCLUSION

This paper introduces a novel local low-rank constrained example based VOC method for estimating the object count which achieves a better performance than existing algorithms even with a complex background. This is because we take advantage of local low-rank constrain to choose samples just from one subspace rather than mixed subspaces at the reconstruction stage, which enhances the performance. Extensive experiments conducted on public datasets validate the effectiveness of our method regardless of the insufficient data or complex background.

ACKNOWLEDGMENT

This work was partially supported by the Shenzhen Science & Technology Fundamental Research Program (No: JCYJ20150430162332418) & (No: JCYJ20160330095814461).

6. REFERENCES

[1] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in Neural Information Processing Systems*, 2010, pp. 1324-1332.

[2] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature Mining for Localised Crowd Counting," in *BMVC*, 2012, p. 3.

[3] Y. Wang, Y. X. Zou, J. Chen, X. L. Huang, and C. Cheng, "Example-based visual object counting with a sparsity constraint," presented at the 2016 IEEE International Conference on Multimedia and Expo (ICME), 2016.

[4] S. An, W. Liu, and S. Venkatesh, "Face recognition using kernel ridge regression," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1-7.

[5] Z. Zhang, M. Wang, and X. Geng, "Crowd counting in public video surveillance by label distribution learning," *Neurocomputing,* 2015.

[6] A. B. Chan and N. Vasconcelos, "Counting people with low-level features and Bayesian regression," *IEEE Transactions on Image Processing,* vol. 21, pp. 2160-2177, 2012.

[7] A. B. Chan and N. Vasconcelos, "Bayesian Poisson regression for crowd counting," in *International Conference on Computer Vision*, 2009, pp. 545-551.

[8] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. López-Sastre, S. Maldonado-Bascón, and D. Oñoro-Rubio, "Extremely Overlapping Vehicle Counting," in *Pattern Recognition and Image Analysis*, ed: Springer, 2015, pp. 423-431.

[9] A. Albiol, A. Albiol, and J. Silla, "Statistical video analysis for crowds counting," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 2569-2572.

[10] M. V. Giuffrida, M. Minervini, S. A. Tsaftaris, and U. Edinburgh, "Learning to Count Leaves in Rosette Plants," 2015.

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

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

[13] Y. Zhou and J. Luo, "A practical method for counting arbitrary target objects in arbitrary scenes," in *2013 IEEE International Conference on Multimedia and Expo (ICME)*, 2013, pp. 1-6.

[14] D. Arpit, G. Srivastava, and Y. Fu, "Locality-constrained low rank coding for face recognition," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, 2012, pp. 1687-1690.

[15] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science,* vol. 290, pp. 2323-2326, 2000.

[16] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055,* 2010.

[17] Z. Ma, L. Yu, and A. B. Chan, "Small Instance Detection by Integer Programming on Object Density Maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3689-3697.