PEOPLE COUNTING WITH IMAGE RETRIEVAL USING COMPRESSED SENSING

Homa Foroughi, Nilanjan Ray, Hong Zhang

Department of Computing Science, University of Alberta, Canada

ABSTRACT

The estimation of the number of people present in an image has many applications such as intelligent transportation, urban planning and crowd surveillance. Rather than conventional counting by detection or regression/machine-learning methods, we propose an image retrieval approach, which uses an image descriptor to estimate the people count. We review the performance of several image descriptors. In addition, we propose a straightforward global image descriptor for image retrieval based on compressed sensing theory. Extensive evaluations on existing crowd analysis benchmark datasets demonstrate the effectiveness of our image retrieval-based approach compared to state-of-the-art regression-based people counting methods.

Index Terms— compressed sensing, people counting, global image descriptor, sparse representation

1. INTRODUCTION

The literature on people counting presents three conceptually different ways to face this task: counting by people detection, counting by feature clustering and counting by regression. In counting by detection [6], [8], [9] each person in the scene is individually detected using a visual object detector, then detected objects are tracked over time; and the number of people is obtained by the number of tracks. In counting by clustering [10],[11] visual features are identified and tracked over time, then the feature trajectories that exhibit coherent motion are clustered to approximate the number of people. In contrast, counting by regression methods [12], [13], [7] aim to learn a direct mapping from global or local low-level features to the number of objects by the use of supervised machine learning techniques and better performance is usually obtained using non-linear regression methods such as Support Vector Regression (SVR), Gaussian Process Regression (GPR) or Bayesian Poison Regression. In [14] an alternative approach is proposed to counting by regression, using pixel-wise density learning and the number of pedestrians in the region of interest is obtained by integrating over the crowd density map. Although there has been a great progress in regression methods in the recent years, the "best" features set and/or kernel function is highly dependent to the crowd density, training set size and crowd segmentation [12]. Thus, learning the regression/density function is still a challenging problem. To avoid the aforementioned difficulty in people counting application, we propose a simple approach based on image retrieval. If we have an annotated large dataset of pedestrians, people counting for a query image can be efficiently performed by retrieving few closest images from the dataset and computing the average of the counts associated with the retrieved images. This is equivalent to the k-nearest neighbor (k-NN) approach, where classification/regression accuracy increases with increasing the training set size. This is so called "big data" approach that has been pursued in computer vision for various tasks before [23]. The contributions of this paper are two-fold, (a) proposing a new method for people counting using image retrieval approach and (b) introducing a simple global image descriptor based on compressed sensing (CS) theory to be employed in this framework. Also, in this paper, we review the performance of several image descriptors in the image retrieval framework.

The CS principle claims that if a signal is sparse, then under certain mild conditions, it can be reconstructed exactly from a small set of random linear measurements using traceable optimization algorithms. Using CS in developing an image descriptor here is motivated by the fact that background subtraction in a video naturally provides a way to make an image, binary and sparse. The theoretical framework of CS was developed by Candes et al [1] and Donoho [2] and since then there was a growing interest in applying this theory to different imaging applications. Han et al [3] proposed an image representation which decomposes the image into dense and sparse components and the concept of compressed sensing was used to code its sparse component. The decomposition scheme was also used by Akbari et al [4] and Jia et al [5] with some improvements. We believe this is the first effort to directly use CS-based descriptor toward image retrieval.

The remainder of the paper is organized as follows: Section 2 gives a brief overview of CS theory. The proposed people counting approach is described in section 3. In section 4, we review some prevailing global image descriptors evaluated in this paper. Finally, experimental results and discussion are presented in section 5.

2. OVERVIEW OF COMPRESSED SENSING

Suppose we have a N-dimensional vector (signal) $f \in R^N$, which could be represented sparsely in a certain domain by the transform matrix Ψ, namely $f = \Psi s$. Clearly, f and s

978-1-4799-2893-4/14/$31.00 ©2014 Crown

are equivalent representations of signal, with f in time or space domain and s in Ψ domain. If there are at most k non-zero entries in this domain, we can say that f is k-sparse. The CS theory guarantees that f could be recovered exactly by taking random measurements much less than N. In order to take the measurements, we first let Φ denote a $\Omega \times N$ matrix with $\Omega \ll N$, then the random non-adaptive measurements are obtained by a linear system. The whole theory is described in equation (1), which is also shown schematically in figure 1.

$$y = \Phi f = \Phi \Psi s = \Theta s \qquad (1)$$

The CS theory says that the signal could be recovered exactly if the number of measurements obeys the condition $\Omega \geq C_M k \log N$, where C_M is a small constant greater than one. The signal can be reconstructed by solving the following convex optimization problem [1]:

$$\hat{f} = arg\ min \|\Psi^T f\|_1 \qquad s.t.\ \ y = \Phi f. \qquad (2)$$

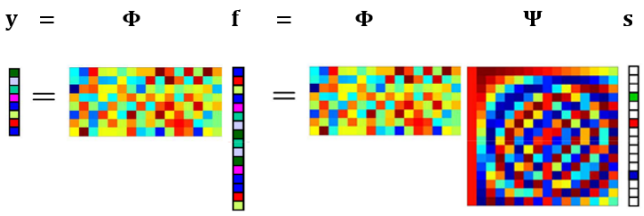


Fig.1. Schematical description of CS theory [15]

3. PROPOSED IMAGE RETRIEVAL APPROACH FOR PEOPLE COUNTING

3.1. The Proposed Method Overview

The goal of proposed method is to estimate the number of people in a given image. Figure 2 gives an overview of our method. Suppose we have large enough pedestrian dataset, which is split into training set used for learning and test set, used for validation. Each image in the dataset is described by an image descriptor and the training phase consists only of storing training set descriptors along with their associated people count. During the test phase, the search engine would retrieve the closest images to the test image using k-nearest neighbor algorithm, according to a similarity measure. To estimate the "people count" for each frame of the test set, its image descriptor would be compared to that of training set, which has been already stored, and estimated count would be the most frequent people count among k closest images to the test frame.

Using the image retrieval framework, we would not be worried about the "best" feature set and regression function; instead we would heavily rely on the raw data by exploiting k-NN algorithm. However, raw image descriptors are very large and highly correlated and this leads to a difficult pattern recognition task. It is observed that k-NN breaks down in high-dimensional space, because of the effect of the curse of dimensionality [25]. To tackle the high dimensional feature space issue, we propose a straightforward image representation based on CS theory, which exploits the sparsity of data, subsequently image descriptors are transformed into a lower dimensional space using PCA.

3.2. CS-based Image Representation

When CS is applied into practical applications such as image representation, there are still some issues to be considered. Natural images are not sparse generally, but compressible in a certain transform domain such as DCT and DWT [3]. Since the goal of this work is to count the number of people in the scene and people are mainly moving objects, so separating out foreground objects from the background, which is called background subtraction, gives the most natural sparse representation of an image for our application. So, in this paper, the sparse representation of image is obtained by Adaptive Gaussian Mixture Model [24] as one the best background subtraction methods.

If Restricted Isometry Property (RIP) satisfies for :

$$(1 - \delta_k)\|x\|_2^2 \leq \|\phi x\|_2^2 \leq (1 + \delta_k)\|x\|_2^2 \qquad (3)$$

one may say, all pairwise distances between k-sparse images are well-preserved in the measurement space. It has been shown, some famous Φs satisfy RIP with high probability like random Gaussian and partial Fourier matrices [28].In this paper, the latter is used in order to exploit translation invariant properties of Fourier transform. So, to take the measurements, k-sparse image (here, binary image) is transformed to frequency domain first.

Clearly, using random Fourier matrices, we are able to exactly reconstruct k-sparse image and approximate compressible image stably with high probability using just Ω random measurements without loss of information [1]. Since these measurements look fair enough to represent an image, our image descriptor is built using the magnitude value of these random measurements. Although CS leads to great compression in the size of image descriptor, in order to avoid high dimensional feature space issues and computational complexity, PCA is applied afterwards to reduce dimension of descriptor. Figure 5 illustrates the error bars of random measurement sampling during several runs.

4. GLOBAL IMAGE DESCRIPTORS

Here, we briefly review some prevailing global image descriptors that will be compared with CS-based image descriptor. Unlike local image descriptors, global descriptors do not require any keypoint detection or matching. They have the ability to generalize an entire image with a single vector, so they are fast to build and efficient to store. As the simplest global image descriptor, we can use the idea of tiny images [23], that images are down-sampled to 64×64 or 32×32 (for grayscale and color images respectively) which is the tolerance of human visual system for scene recognition task. Gabor-Gist [16], which is the most common global descriptor, represents an

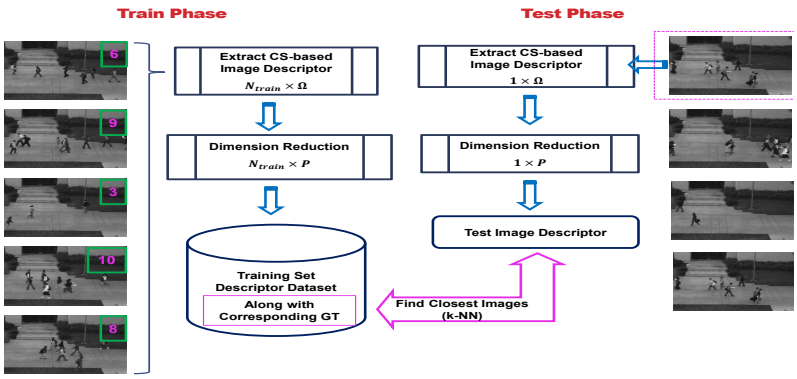
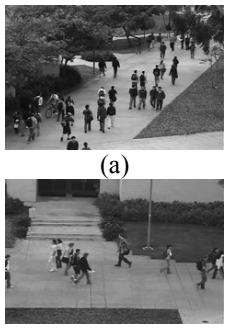


Fig.2. The proposed method overview



(a) (b)

Fig.3. UCSD dataset
(a) Peds1 and (b) Peds2

image in terms of its responses to a bank of Gabor filters. Image is divided into 4×4 image tiles and the final feature descriptor would be the mean response of tiles to steerable filters at different scales and orientations. HOG [17] counts the occurrences of gradient orientation in localized portions of an image. It is a window based descriptor densely sampled over all image points. The window is divided into a square grid and the distribution of edge orientations within each cell is computed. WI-SIFT (Whole Image SIFT) and WI-SURF are the global versions of the corresponding famous local normalized descriptors SIFT [18] and SURF [19]. In this case, the center of image is considered as the only detected keypoint and the image descriptor, describes its neighborhood including whole image. Recently, some local binary descriptors have been presented such as BRIEF [21] and BRISK [22] which are mostly similar in nature and based on comparison of neighboring pixel intensities. BRIEF-Gist was proposed by [20] as a global version of BRIEF, in which rather than extracting local descriptors, the original image is divided into $n \times n$ tiles, a descriptor is built for each patch tile, and the final descriptor would be their concatenation.

5. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed method, we carry out a series of experiments on two pedestrian datasets. The UCSD dataset [12] was collected from two viewpoints overlooking a pedestrian walkway. The first viewpoint (Peds1, figure 3-a) is an oblique view including around 33000 frames containing large number of people (0~46) and the second one (Peds2, figure 3-b) is a side-view including 34000 frames (0~15 people). The first 4000 frames of each video sequence were used for ground-truth (GT) annotation.

Since our model should be validated on large dataset, more frames with ground-truth have to be prepared. Hence, we annotated the whole dataset on the same region of interest and we did our best to be consistent with the provider's team unwritten rules while counting individuals. However, in the future work, we will use Semi Supervised Learning [27] to annotate the large amount of un-labelled data. Since the effectiveness of the proposed method must be validated on a large dataset, FUDAN [26] or Mall [7] datasets would not be useful because they just include up to 2000 frames.

The dataset is split into a training set, for learning and a test set, for validation. In order to review the effect of varying training set size, experiments are conducted with three different training sizes; the first 15k, 20k and 25k frames of each dataset are selected for training and the rest for test purposes. Figure 4 shows three different sample test images; belonging to low, medium and high density scenes as well as top three retrieved image using k-NN algorithm. As it can be seen, the retrieved images of each column, look pretty similar to the test image in terms of people count.

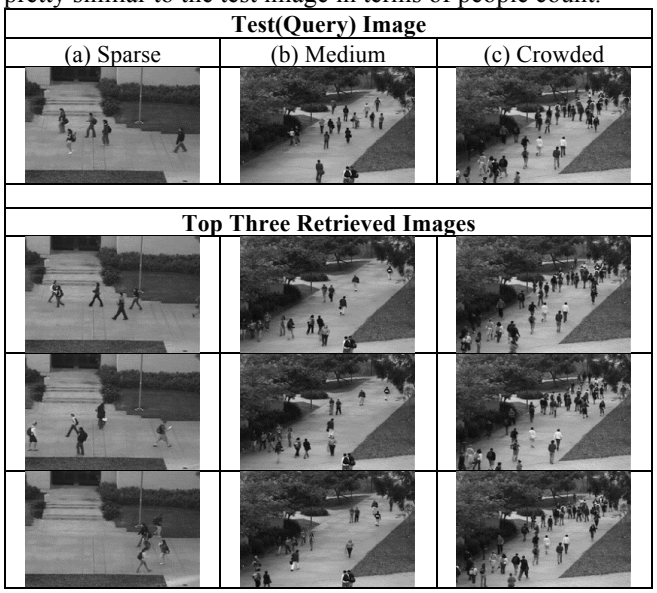


Fig.4. Example of system output

We compare the performance of CS-based proposed descriptor with other aforementioned global descriptors. The accuracy of estimates is evaluated by the mean square error and mean absolute error as follows:

$$MSE = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (\hat{c}_i - c_i)^2 \qquad MAE = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} |\hat{c}_i - c_i|$$

where c_i and $\hat{c}_i$ are the true and estimated count for frame i.

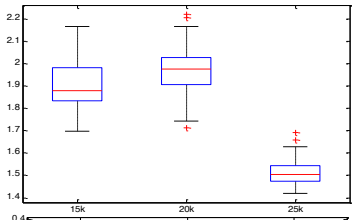
Table I. Results on Peds1 dataset

Method	MAE Train Size			MSE Train Size		
	15k	*20k*	*25k*	*15k*	*20k*	*25k*
CS-based-PCA	1.900	1.868	1.517	6.384	5.039	4.628
Gist	2.262	2.172	1.760	9.123	8.515	5.643
Gist-PCA	2.274	2.162	1.769	9.267	8.504	5.708
HOG	1.714	1.639	1.373	5.502	4.917	3.546
WI-SIFT	2.322	2.134	1.667	10.257	8.128	5.075
WI-SURF	2.963	2.901	2.406	16.499	15.291	10.941
BRIEF-Gist	4.546	4.157	3.061	32.567	28.009	18.007
SubSample	3.492	3.423	3.094	19.737	18.819	16.198
SubSample-PCA	4.503	4.286	3.014	31.237	30.008	17.683
GPR	2.985	2.999	3.421	13.084	13.357	15.904
SVR	2.873	2.119	2.217	12.328	7.065	7.145

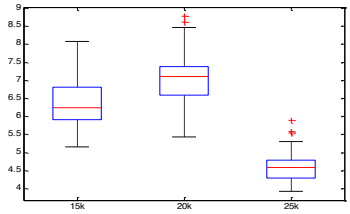
Table II. Results on Peds2 dataset

Method	MAE Train Size			MSE Train Size		
	15k	*20k*	*25k*	*15k*	*20k*	*25k*
CS-based-PCA	0.343	0.285	0.173	0.541	0.430	0.226
Gist	0.372	0.311	0.175	0.608	0.507	0.239
Gist-PCA	0.382	0.318	0.184	0.612	0.515	0.251
HOG	0.369	0.325	0.194	0.553	0.479	0.257
WI-SIFT	0.364	0.319	0.194	0.583	0.528	0.257
WI-SURF	0.474	0.426	0.244	0.851	0.761	0.352
BRIEF-Gist	1.287	1.194	0.778	3.697	2.811	1.387
SubSample	1.203	0.890	1.197	2.681	1.423	2.269
SubSample-PCA	0.691	0.597	0.307	1.618	1.308	0.477
GPR	4.466	4.619	4.915	20.858	22.293	25.048
SVR	1.368	1.465	1.695	3.125	3.394	4.381

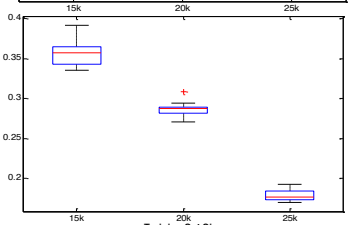
(a) MAE error bar on Peds1



(b) MSE error bar on Peds1



(c) MAE error bar on Peds2



(d) MSE error bar on Peds2

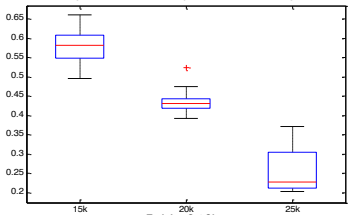


Fig.5. MAE and MSE error bars of random measurement sampling of CS-based descriptor on 30 runs using different tarining set size

The results on Peds1 and Peds2 datasets are summarized in tables I and II. The similarity between image descriptors is calculated using Euclidean distance except for BRIEF-Gist and sub-sampling methods, which Hamming distance and sum of squared differences [23] is replaced. We also compare our approach with two state-of-the-art counting by regression methods. In these methods, firstly three categories of features including segments, internal edges and texture are extracted form images [12] and for the regression function, GPR [12] or SVR is employed. For implementations, we used GPML [29] and LIBSVM [30].

According to the results, a number of conclusions are possible. First, the image retrieval method has much superior performance than the regression methods. All of the global image descriptors including CS-based, outperform both GPR and SVR methods. Second, as the training set size increases, both MAE and MSE are decreased for all the descriptors, which proves the effectiveness of this framework for large datasets. However, in the case of regression methods, involving more data is not always helpful. By increasing the training size, the degree of non-linearity can be increased in the chosen feature space, and the chosen kernel function might not be able to capture it well. This justifies the dramatic drop of performance in GPR, however in SVR, no clear trend is inferred; after a decrease, errors are increased afterwards. Third, we observe that the best overall performance is achieved by CS-based image descriptor in Peds2 dataset; however crowded images in Peds1 are better modeled with HOG descriptor and our descriptor stands in the second place. Peds1 contains larger crowds and it is more potential for errors because of traveling bicycles, skateboarders and golf carts. In future work, the full power of CS-based descriptor will be exploited by adaptively choosing one random subset out of many possible random measurement options.

6. CONCLUSION

In this paper, we proposed a novel approach for people counting based on image retrieval which uses an image descriptor to estimate the count. In addition to reviewing the performance of prevailing global descriptors, we introduced a compact global image descriptor based on CS theory. The experimental results reveal that proposed approach performs well to estimate crowd density in comparison with state-of-the-art regression methods. We validated our method on two challenging datasets. The advantage of this approach is particularly significant when the pedestrian dataset is large.

7. REFERENCES

[1] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," Information Theory, IEEE Trans., vol. 52, no. 2, pp. 489-509, Feb.2006.

[2] D. Donoho, "Compressed sensing", Information Theory, IEEE Trans., vol. 52, no. 4, pp. 1289-1306, Feb.2006.

[3] Bing Han, Feng Wu and Dapeng Wu, "Image representation by compressed sensing" Image Processing, In 15th IEEE International Conference on 2008, Page(s): 1344 – 1347

[4] A. Akbari, P. Zadeh, M. Moniri, "Stereo image representation using compressive sensing" in 3DTV Conference: The True Vision Capture, Transmission and Display of 3D Video (3DTV-CON), 2011, May 2011,pp. 1–4.

[5] Yingbiao Jia; Yan Feng; Yuming Cao; Changsheng Dou, "The algorithm of image representation and reconstruction based on compressed sensing with composite measurement," Communication Software and Networks (ICCSN), pp.404,407, May 2011

[6] M. Li, Z. Zhang, K. Huang, and T. Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In International Conference on Pattern Recognition, pages 1–4, 2008.

[7] K. Chen, C.C. Loy, S. Gong, T. Xiang, "Feature mining for localised crowd counting", in: Proceedings of British Machine Vision Conference, 2012, pp. 21.1–21.11.

[8] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 30, no. 7, pp. 1198–1211, 2008.

[9] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 875–85

[10] G. J. Brostow and R. Cipolla. Unsupervised Bayesian detection of independent motion in crowds. In IEEE Conference on Computer Vision and Pattern Recognition, pages 594–601, 2006.

[11] V. Rabaud and S. J. Belongie, "Counting crowded moving objects," in IEEE Conf. Computer Vision and Pattern Recognition, 2006.

[12] A.B. Chan and N. Vasconcelos. Counting people with low-level features and Bayesian regression. IEEE Transactions on Image Processing, 21(4):2160–2177, 2012.

[13] D. Kong, D. Gray, and H. Tao. Counting pedestrians in crowds using viewpoint invariant training. In British Machine Vision Conference, 2005.

[14] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in Advances in Neural Information Processing Systems, 2010.

[15] Richard Baraniuk, Justin Romberg, and Michael Wakin, Tutorial on compressive sensing (2008 Information Theory and Applications Workshop)

[16] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope", International Journal of Computer Vision, 42(3), pp. 145-175, 2001.

[17] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, pp. 886-893, 2005.

[18] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, 60(2), pp. 91-110, 2004.

[19] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded-up robust features. Computer Vision and Image Understanding, 110(3):346–359, June 2008.

[20] N. S¨underhauf and P. Protzel, "BRIEF-Gist – Closing the Loop by Simple Means", IEEE/RSJ International Conference on Intelligent Robots and Systems, USA, pp. 1234-1241, 2011.

[21] M. Calonder, V. Lepetit, C. Strecha, et al., "BRIEF: Binary Robust Independent Elementary Features, European Conference on Computer Vision, Crete, Greece, pp. 778-792, 2010.

[22] S. Leutenegger, M. Chli and R. Y. Siegwart, BRISK: "Binary Robust Invariant Scalable Keypoints", IEEE International Conference on Computer Vision, Spain, pp. 2548-2555, 2011.

[23] A. Torralba, R. Fergus, and W. Freeman. "80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11):1958–1970, 2008.

[24] Z. Zivkovic, "Improved Adaptive Gaussian Mixture Model for Background Subtraction," Proc. Int'l Conf. Pattern Recognition, vol. 2, pp. 28-31, 2004.

[25] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is nearest neighbor meaningful," in Proc. 7th Int. Conf. Database Theory, 1999, pp. 217 – 235

[26] FUDAN Dataset: http://www.iipl.fudan.edu.cn/~zhangjp/Dataset/fd_pede_dataset_intro.htm

[27] Chapelle, O., Zien, A., and Scholkopf, B. (Eds.). (2006c). Semi-supervised learning. MIT Press.

[28] Fan Yang; Shengqian-Wang; Chengzhi Deng, "Compressive sensing of image reconstruction using multi-wavelet transforms," ICIS, 2010, vol.1, no., pp.702,705, 29-31

[29] http://www.gaussianprocess.org/gpml/code/matlab/doc/

[30] http://www.csie.ntu.edu.tw/~cjlin/libsvm/