Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Robust people counting using sparse repregcisentation and random projection

Homa Foroughi *, Nilanjan Ray [1], Hong Zhang [2]

2-21 Athabasca Hall, Department of Computing Science, University of Alberta, Edmonton, AB, Canada, T6G 2E8

ARTICLE INFO

Article history:
Received 29 September 2014
Received in revised form
16 January 2015
Accepted 13 February 2015

Keywords:
People counting
Sparse representation
Fast l_1−minimization
Random projection
Convolutional neural network
Semi-supervised learning

ABSTRACT

Estimating the number of people present in an image has many practical applications including visual surveillance and public resource management. Recently, regression-based methods for people counting have gained considerable importance, principally due to the capability of these methods to handle crowded scenes. However, the principal drawback of regression-based methods is to find an optimal set of features and a model, which is usually dependent on the crowd density. Encouraged by the recent success of sparse representation, here, we develop a robust and scalable people counting method. Sparse representation allows us to capture the hidden structure and semantic information in visual data and leads to faster processing algorithms. In order to reduce the complexity of solving l_1−minimization problem, which resides at the heart of the sparse representation, a dimensionality reduction method based on random projection is employed. The sparse representation framework provides new insight that if sparsity in the classification problem is properly harnessed, feature extraction is no longer critical. So, in addition to several hand-crafted features, we exploit the features obtained from pre-trained deep Convolutional neural network and show these features perform competitively. Further, to render the proposed method user friendly, we employ a semi-supervised elastic net to automatically annotate unlabelled data with only a handful of user-labelled image frames. Our semi-supervised method exploits temporal continuity in videos. We use extensive evaluations on the crowd analysis benchmark datasets to demonstrate the effectiveness of our approach as well as its superiority over the state-of-the-art regression-based people counting methods, in terms of accuracy and time.

1. Introduction

Estimation of the number of people in a scene is a topic of significant interest in areas such as safety and security, resource management, urban planning and scheduling public transportation systems. Literature on people counting includes three conceptually different techniques: counting by people detection, counting by clustering and counting by regression.

In the counting by a detection technique [1,2], a classifier is trained using the common features of pedestrian training images, which usually include Haar-wavelets or histogram of oriented gradients (HOG) [3]. A trained classifier is then applied in a sliding window fashion across the whole image space to detect pedestrian candidates. The detection performance can be further improved by adopting a part-based detection technique or tracking validation during frames. But, as the crowd becomes larger and denser, detection and tracking tasks become impractical due to occlusions. An alternative way is counting by clustering [4,5] which consists of the steps of identifying and tracking visual features over time. This technique assumes a crowd to be composed of individual entities, each of which has a unique yet coherent motion pattern that can be clustered to estimate the number of people. However, it needs sophisticated trajectory management and in crowded environments, coherently moving features usually do not belong to the same person. The counting by the regression technique [6,7] counts people by learning a direct mapping from low-level image features to the number of people by the use of supervised machine learning algorithms. A popular approach is to extract several global features with complementary nature from crowd segments and combine them to form a bank of features and then a regression function is trained to predict the people count. This technique avoids segmentation/detection of individuals and estimates the crowd density based on a holistic and collective description of crowd patterns. Although counting by regression is feasible for crowded environments and could achieve promising results, it still suffers from serious weaknesses. In particular, Loy et al. [8] reveal that the optimal feature set is different in sparse and crowded scenes. In fact, the number of features carried by one

* Corresponding author. Tel.: +1 780 492 6365, fax: +1 780 492 6393.
E-mail addresses: homa@ualberta.ca (H. Foroughi), nray1@cs.ualberta.ca (N. Ray), zhang@cs.ualberta.ca (H. Zhang).
[1] Tel.: +1 780 492 3010, fax: +1 780 492 1071.
[2] Tel.: +1 780 492 7188, fax: +1 780 492 1071.

http://dx.doi.org/10.1016/j.patcog.2015.02.009
0031-3203/

pedestrian is heavily affected by camera perspective and crowd density, also it is observed that different features can be more important given various crowdedness levels. In addition, their evaluations show that the actual performance of a regression model can be quite different from what one may anticipate, subject to the nature of data, especially when it is applied to unseen crowd density.

Unlike regression techniques, our proposed method based on sparse representation, does not need to select either the optimal feature set or the regression model. The main idea behind sparse representation is, if a collection of representative samples are found, we should expect that a typical sample has a very sparse representation with respect to such a learned basis. In other words, given sufficient diversity in the training images, the new test image can be well represented as a sparse linear combination of the training set. This sparse representation would naturally encode the semantic information of the image [9]. In order to reduce the time complexity of finding the sparse representation, random projection is utilized as our choice of dimensionality reduction method.

It is commonly believed that the Sparse Representation-based Classification (SRC) requires a rich set of training images of every class that can span the variation under testing conditions. To fulfill this requirement, we use a semi-supervised learning framework to avoid exhaustive manual image annotation. Extensive experimental results suggest that our proposed method is fast, accurate and scalable to large-scale datasets.

The remainder of the paper is organized as follows: the theory of sparse representation is summarized in Section 2. Section 3 shows how to apply general classification framework to people counting task. In Section 4 we discuss how we exploit semi-supervised regression to deal with few labelled training samples effectively. Experimental setup is explained in Section 5 and results and discussion are presented in Section 6, followed by conclusion remarks in Section 7.

2. Sparse representation

Sparse representation (SR) has proven to be an extremely powerful tool for acquiring, representing, and compressing high-dimensional signals. This success is mainly due to the fact that important classes of signals such as audio and images have naturally sparse representations with respect to fixed bases e.g. Fourier and Wavelet. Moreover, in recent years, efficient and fast algorithms have been proposed for computing such representations [9]. The problem solved by sparse representation is to search for the most compact representation of a signal (image) in terms of a linear combination of relatively few base elements in a basis or over-complete dictionary. If the optimal representation is sufficiently sparse, it can be efficiently computed by greedy methods or convex optimization. Typically, the sparse representation technique is cast into an l_1-minimization problem, which is equivalent to the l_0-minimization under some conditions. This l_0-l_1 equivalence has provided computational convenience as evidenced by Compressed Sensing (CS) [10].

In the recent years, variations and extensions of l_1-minimization have been applied to many computer vision tasks, including face recognition [11], background modelling [12] and image classification [13]. In almost all of these applications, using sparsity as a prior leads to the state-of-the-art results [9]. The ability of sparse representation to uncover semantic information derives in part from a simple but important property of the data: although the images (or their features) are naturally very high dimensional, in many applications images belonging to the same class exhibit degenerate structure. That is, they lie on or near low-dimensional subspaces or submanifolds [9]. So, if a collection of representative samples are found, we should expect that a typical sample has a very sparse representation with respect to such a (possibly learned) basis. Such a sparse representation, if computed correctly, could naturally encode the semantic information of the image [9]. SRC seeks a sparse representation of the query image in terms of the over-complete dictionary and then performs the recognition by checking which class yields the least representation error. SRC can be considered as a generalization of Nearest Neighbor (NN) and Nearest Feature Subspace (NFS). Generally speaking, Nearest Feature based Classifiers (NFCs) aim to find a representation of the query image, and classify it to the best representor. According to the mechanism of representing the query image, NFCs include Nearest Neighbor, Nearest Feature Line (NFL), Nearest Feature Plane (NFP) and Nearest Feature Subspace. More specifically, NN is the simplest one with no parameters, which classifies the query image to its nearest neighbor. NN, NFL and NFP all use a subset of the training samples with the same label to represent the query image, while NFS represents the query image by all the training samples of the same class. In general, the larger samples lead to better stability of a method. The most generalized classifier is SRC, which considers all possible supports (within each class or across multiple classes) and adaptively chooses the minimal number of training samples needed to represent each test sample. In the next section, we show how this sparse representation can be used in people counting application.

3. People counting based on sparse representation and random projection

3.1. People counting as sparse representation

Suppose that we have a set of labelled (annotated) training images from a pedestrian dataset where the number of people present in each image is given. We assume these labelled training images $\{x_i, l_i\}$ are from C different classes. Here, class (label) l_i is equal to the count, i.e. number of people in the image $x_i \in R^m$, where x_i is the vector representation of the image, which could be its raw pixels or features computed from the raw pixels. Given sufficient training samples from the ith class, any new test sample $x_{test} \in R^m$ from the same class, will approximately lie in the linear span of the training samples associated with class i.

$$x_{test} \approx \Sigma_{\{j|\, l_j = i\}} x_j \alpha_j = X_i \alpha_i \tag{1}$$

where $X_i \in R^{m \times n_i}$ concatenates all of the images of class i. Since the class label of the test image is initially unknown, we would form a linear representation similar to Eq. (1), now in terms of all training samples. We define a new matrix (dictionary) $\Psi \in R^{m \times n}$ for the entire training set as the concatenation of all $n = \Sigma_i n_i$ training samples of all C classes:

$$x_{test} = [X_1, X_2, \ldots, X_C]\alpha = \Psi\alpha \in R^m \tag{2}$$

where

$$\alpha = [\ldots, 0^T, \alpha_i^T, 0^T, \ldots]^T \in R^n \tag{3}$$

α is a coefficient vector whose entries are zero except those associated with the ith class. We notice that α is a highly sparse vector and on average, only a fraction of $1/C$ coefficients are nonzero and the dominant nonzero coefficients in the sparse representation α reveal the true class of test image. Indeed, in the test phase, we wish to represent a new unlabelled image in a Ψ-dependent space in which the image has a sparse representation. In general, this vector is the sparsest solution to the system of equations $x_{test} = \Psi\alpha$ which is found by solving the following optimization problem:

$$\alpha^* = \text{argmin}\|\alpha\|_0 \quad \text{s.t. } \Psi\alpha = x_{test} \tag{4}$$

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

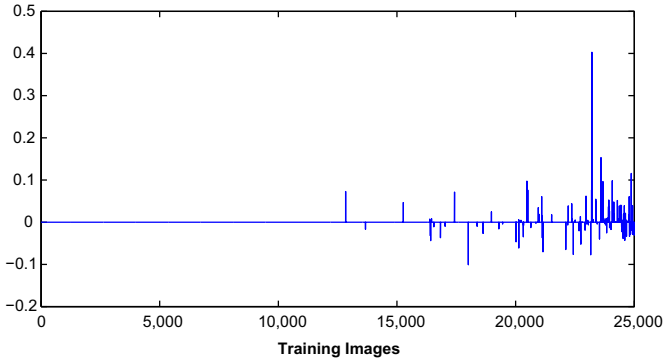

Fig. 1. Representation of a test image as sparse linear combination of the training set.

While the search for a sparse solution to a linear system is a difficult problem in general, foundational results in the theory of sparse representation show that the sparsest solution can be exactly recovered by solving a tractable optimization problem [14]. This is achieved by seeking α as the unique solution to the optimization problem:

$$\alpha^* = \text{argmin}\|\alpha\|_1 \quad \text{s.t. } \|\Psi\alpha - x_{test}\|_2 \le \epsilon \tag{5}$$

where ϵ is the noise level in the observation. Eq. (5) can be rewritten as follows using a Lagrangian multiplier:

$$\alpha^* = \text{argmin}\|\Psi\alpha - x_{test}\|_2^2 + \lambda\|\alpha\|_1 \tag{6}$$

Fig. 1 shows an example of vector α recovered by solving Eq. (6).

3.2. The role of feature extraction

Conventional visual representations for creating the dictionary Ψ, include local and global image descriptors. Local descriptors like SIFT [15] are found on the premise that images can be characterized by attributes computed on regions of the image; however, global (whole image) descriptors do not require any keypoint detection. They have the ability to characterize an entire image with a single vector. Also, they are fast to build and efficient to store. Moreover, the previous studies [16] show that global descriptors are highly effective in describing the crowd density when enough training samples are involved. Gist [17] represents an image in terms of its responses to a bank of Gabor filters with different frequencies and orientations. An image is divided into tiles and the final feature descriptor would be the mean response of tiles to steerable filters. HOG [3] 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.

Recently, in contrast to the hand-crafted features, learnt image features with deep network structures have shown great potential in various vision recognition tasks. Among these architectures, one of the greatest breakthroughs in image classification is the deep Convolutional Neural Network (CNN) [18], which has achieved the state-of-the-art performance in the large-scale object recognition task. Thanks to an optimized GPU implementation and new regularization techniques, Krizhevsky et al. [18] successfully applied CNNs to classification on the ImageNet dataset [19]. The feature representation learned by this network shows excellent performance not only on the ImageNet classification task it was trained for, but also on a variety of other recognition tasks [20,21]. In this paper, we would evaluate whether features extracted from a pre-trained CNN can be reused to the people counting task. To the best of our knowledge, this is the first attempt to employ deep learning to solve the pedestrian counting problem. For training CNNs and subsequent feature extraction, we use the DeCAF code [20] and instead of training a network from scratch, we use a pre-trained model on ImageNet. We refer the reader to [18] for more details on the architecture and training algorithm, which we followed exactly. As the feature descriptor, we use $DeCAF_7$; the features taken from the final hidden layer, i.e. just before propagating through the final fully connected layer to produce the class predictions.

3.3. Dimensionality reduction

One major obstacle in real-world large-scale people counting is the large scale of the training data and the high dimensional feature vectors. A classical technique for addressing the problem of high dimensionality is to project the data into a much lower-dimensional feature space $R^d (d \ll m)$, such that the projected data still retain the useful properties of the original images. From Matten et al.'s study [22] one may infer that, all non-linear dimensionality reduction techniques require the optimization of some parameters and some of them suffer from memory complexity issue. Also, when $m < n$, where m is the feature dimension and n is the number of training data, non-linear techniques have computational disadvantages compared to the classical linear algorithm, such as principal component analysis (PCA) [23]. Thus, we selected PCA as the baseline dimensionality reduction technique for our application; however, PCA is still expensive for high-dimensional and large-scale image data; the computational complexity of PCA is $O(m^2n + m^3)$ due to the matrix computation and SVD eigenvalue decomposition [24]. A statistically optimal way of dimensionality reduction is to project the data onto a random lower-dimensional subspace that captures as much of the variation of data as possible. This technique is called Random Projection (RP), which is computationally efficient and simple. In RP, the original m-dimensional data is projected to a d-dimensional subspace through the origin, using a random matrix $\Phi_{d\times m}$ whose columns have unit lengths. The key idea of random projection arises from the Johnson–Lindenstrauss lemma [25]: if points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved. In contrast, PCA does not guarantee (approximate) distance preservation between data point pairs. Random projection is computationally very simple; its complexity is bounded by $O(mdn)$ and this could be even less when the data matrix is sparse [24]. Applying such a linear projection on training and test images, gives us a new observation:

$$\tilde{x}_{test} = \Phi x_{test} \in R^d$$
$$\tilde{\Psi} = \Phi\Psi \in R^{d\times n} \tag{7}$$

Then the sparsest solution to α, would be obtained via a lower complexity convex optimization:

$$\alpha^* = \text{argmin}\|\tilde{\Psi}\alpha - \tilde{x}_{test}\|_2^2 + \lambda\|\alpha\|_1 \tag{8}$$

Since $d \ll m$, the complexity is significantly reduced. RP offers clear benefits over PCA; the choice of projection does not depend on the data, it is much faster and as the projected dimension is decreased and drops below a threshold, RP offers a gradual degradation in the performance; however, the degradation suffered by PCA is not necessarily so gradual. At smaller dimensions, PCA distorts the data and this is mainly because the performance of PCA is dependent on the sum of omitted eigenvalues [24]. The choice of random matrix Φ is one of the key points of interest. The elements of Φ are often Gaussian distributed, i.e. the entries are independently sampled from a zero mean normal distribution $N(0, 1)$ and each row is normalized to unit length (makes rows orthogonal),

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

and this is usually called Gaussian random matrix [11]. Another eligible family is partial Fourier, or more generally, random rows of any orthonormal matrix.

3.4. Count estimation

Once the sparse vector α^* has been recovered by Eq. (8), we can estimate the count for the unlabelled test image. This step can be regarded as class recognition in usual object classification approaches. Ideally nonzero entries in α^* will all be associated with single class and we can easily assign the test sample to that class. However, noise and modelling error may lead to small nonzero entries associated with multiple (other) classes. One may design different classifiers to resolve this issue. For instance, we can simply assign x_{test} to the class with the largest (or few largest) entry in α^*. However, such heuristics do not harness the subspace structure associated with images [11]. An alternative way is assigning x_{test} to the class with the most contribution in α^* i.e. the one with the most number of nonzero coefficients in sparse vector α^*. We know that $\alpha^* = [\alpha_1^T, \alpha_2^T, \ldots, \alpha_C^T]^T$, then we would compute the following ratio for each class i:

$$r_i = \frac{\|\alpha_i\|_1}{\|\alpha\|_1} \tag{9}$$

And then assign x_{test} to the label i that maximizes r_i [14]. A slightly more sophisticated schema is, we classify x_{test} based on how well the coefficients associated with all training samples of each class reproduce x_{test}. We do it by comparing how well the different parts of the estimated coefficients α^* represent x_{test} and the minimum of the representation error or the residual error is then used to identify the correct class [11]. For each class i, let δ_i as a function that selects the coefficients associated with ith class. For $\alpha \in R^n$, $\delta_i(\alpha) \in R^n$ is a vector whose only nonzero entries are associated with class i. Using only the coefficients associated with the ith class, one can approximate the given test sample x_{test} as $\widehat{x}_{test} = \Psi\delta_i(\alpha^*)$. We then classify x_{test} to the class that minimizes the residual error between x_{test} and $\widehat{x}_{test}$:

$$\min_i \mathrm{res}_i(x_{test}) = \|x_{test} - \delta_i(\alpha^*)\|_2 \tag{10}$$

3.5. Sparse solution via fast l_1-minimization

Although dimensionality reduction reduces the time complexity significantly, finding an efficient and fast way to solve Eq. (8) would be extremely beneficial. While the l_1-minimization problems associated with CS can be formulated as a linear program and readily solved by classical methods in convex optimization, such as interior-point methods [26,27], the computational complexity of those classical second-order methods is often too high for high-dimensional image data. These methods are accurate but problematic because of their need to solve large systems of linear equations to compute Newton steps and suffer from poor scalability for large-scale real-world problems such as people counting. Due to a large number of real applications in various fields, many new efficient accelerated algorithms have been proposed over the past decade. In this section we describe just our selected algorithms for solving the optimization problem for the people counting application and we draw extensively on the survey [28], which compares the performance of different l_1-minimizers in the context of face recognition. Homotopy is a fast l_1-minimization method for sparse optimization. This approach was first studied in the context of lasso, which inspired a solution to the forward stagewise linear regression problem called LARS [29] and eventually led to the Homotopy algorithms for basis pursuit in [30,31]. Donoho et al. [31] showed that the Homotopy runs much more rapidly than general-purpose solvers when sufficient sparsity is present. If the underlying solution has only d nonzeros, the Homotopy method reaches that solution in only d iterative steps. The computational cost of the Homotopy for l_1-minimization is bounded by $O(dm^2 + dmn)$ and this is a significant improvement from the interior-point methods. According to the computational cost and extensive experiments provided in [28], although the Homotopy is not a fast method for recovering non-sparse signals, it performs very fast and accurate when the signal is extremely sparse.

Inspired by iterative thresholding ideas, a great number of first-order methods are also available. These methods are more efficient and more scalable which can solve very large-scale l_1-minimization problems to medium accuracy. A special class of first-order methods is called Augmented Lagrangian Methods (ALM) [32] in which their basic idea is to eliminate equality constraints by adding an appropriate penalty term to the cost function that assigns a very high cost to infeasible points. The goal is then to efficiently solve the unconstrained problem using an accelerated gradient algorithm [14]. ALM methods differ from other penalty-based approaches by simultaneously estimating the optimal solution and Lagrange multipliers iteratively. Compared to the classical interior-point methods, ALM generally takes more iterations to converge to the optimal solution; however, the biggest advantage of ALM is that each iteration is composed of very elementary matrix–vector operations as against matrix inversions or Gaussian eliminations used in the interior-point methods. The principles of ALM can also be applied on the dual problem of Eq. (8) and according to comprehensive comparisons provided in [28], the dual ALM (DALM) algorithm performs the best in the face recognition experiment, and scales well in terms of the number of subjects. Computational time of DALM is dominated by $O(m^2 + mn)$ that makes it one of the best l_1-minimization methods [28]. In brief, the Homotopy and DALM solvers provide a viable solution to real-world; time-critical applications such as people counting in which speed and scalability both play an important role.

4. Undersampled SRC

SRC shows powerful discriminative power when the training samples are sufficient to construct an over-complete dictionary. It is commonly believed that SRC always requires a rich set of training images of each class that can span variations of that class under testing conditions; however, without adequate training samples as it is common in the real-world applications, the dictionary is too small to sparsely represent the test sample, yielding poor classification performance [33]. If the dictionary consists of few images from each class, the nonzero coefficients of α^* would not concentrate on the correct class. This happens when we cannot afford to obtain sufficient training samples (e.g biometric datasets) or either when the collected training samples are unlabelled (e.g. pedestrian datasets). For the former, some adaptive SRC methods have been proposed recently to deal with a small dictionary using an auxiliary interclass variant dictionary to represent the possible variations between the training and test images for face recognition task [34] and constructing a graph Laplacian which encodes the sparse representation relationship of all samples [33]. In the case of latter, obviously increasing the quantity and diversity of hand-labelled images improves the performance of SRC; however, manual image annotation is a tedious and labor-intensive task. Fortunately, in pedestrian datasets, there are a large number of unlabelled frames which provide useful topological information. To exploit such an intrinsic distribution structure, we develop a semi-supervised learning framework; in which instead of exhaustively annotating every single frame, only a handful of frames are selected for annotation automatically and actively. Here, the goal is to provide sufficient labelled training samples for SRC via automatic image annotation, which is basically labelling images with number of people in them

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

in our application. We take an approach similar to [35] which uses Semi-Supervised Elastic Net (SSEN) regression to predict frame count, by incorporating correlation of sequential unlabelled frames to penalize sudden prediction change. We then develop an iterative procedure in which SSEN predicts the counts on the unlabelled images and then the most confident unlabelled samples together with their predicted labels are added to enlarge the training set. The labelling procedure would be repeated until sufficient training samples are provided.

4.1. Elastic net

Elastic net (EN) is a novel shrinkage and selection method for producing a sparse model with good prediction accuracy. It encourages the grouping effect and elegantly handles the high-dimensionality by reducing the redundant features which have less relationship with the properties if pedestrians. Given a set of l labelled samples $\{(x_i, y_i)\}_{i=1}^{l}$, where $x_i \in R^m$ is the vector representation of the image and y_i is the corresponding label, EN optimizes the following function:

$$\hat{\beta} = \text{argmin}\|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2 \tag{11}$$

where $X \in R^{n\times m}$ is a stacked matrix of x_is, $y \in R^{n\times 1}$ is the concatenation of y_is and λ_1 and λ_2 are tuning parameters. To achieve the solution of EN, let

$$X^*_{(n+m)\times m} = (1+\lambda_2)^{-1/2}\begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix} \quad \text{and} \quad y^*_{(n+m)\times 1} = \begin{pmatrix} y \\ 0 \end{pmatrix}$$

then Eq. (11) can be rewritten:

$$\hat{\beta^*} = \text{argmin}\|y^* - X^*\beta^*\|_2^2 + \gamma\|\beta^*\|_1 \tag{12}$$

where $\gamma = \lambda_1/\sqrt{(1+\lambda_2)}$; the detailed transformations can be found in [36].

4.2. Semi-supervised elastic net

One common way to construct a semi-supervised algorithm is to add unlabelled data as a regularization term. In SSEN, this term is created by exploiting the sequential information among unlabelled frames. More precisely, given a set of training data, we assume some of them are labelled, $L = \{(x_i, y_i)\}_{i=1}^{l}$, but most of them are unlabelled, $U = \{x_j\}_{j=l+1}^{l+u}$. The user would only label a few data points and the rest of unlabelled training data will be annotated automatically using the semi-supervised model. The goal is to learn a function from both small minority L and vast majority U. Generally, we observe that in most sequential frames, pedestrian quantities change slightly or even remain unchanged. Tan et al. [35] used this fact to add a regularization term to EN in order to penalize sudden prediction change between neighboring frames. So we would have

$$\hat{\beta} = \text{argmin}\|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2 + \lambda_3 B \tag{13}$$

where $B = \|D\beta\|_2^2$ and D is a matrix that each row is obtained by difference of x_is of neighboring frames. Like before, Eq. (13) can be simplified as

$$\hat{\beta^*} = \text{argmin}\|\tilde{y} - \tilde{X}\beta^*\|_2^2 + \gamma\|\beta^*\|_1 \tag{14}$$

where

$$\tilde{X}_{(n+m+\|\Omega\|)\times m} = (1+\lambda_2)^{-1/2}\begin{pmatrix} X \\ \sqrt{\lambda_2} I \\ \sqrt{\lambda_3} D \end{pmatrix} \quad \text{and}$$

$$\tilde{y}_{(n+m+\|\Omega\|)\times 1} = \begin{pmatrix} y \\ 0 \\ 0 \end{pmatrix}$$

where $\|\Omega\|$ denotes the cardinality of neighboring frame set Ω which is constructed from neighboring unlabelled frames. Eq. (14) is a lasso-type function that is usually solved using the LARS method [29]; however it can be solved much faster and more efficiently by Homotopy or DALM methods descried earlier. The problem of optimally selecting a handful of labelled samples plays a key role in improving annotation performance and saving human-labor. It is believed that given a fixed number of labelling budget, the most representative frames (in the sense of covering different crowd densities/counts) are the most useful ones to label. This brings in a chicken-and-egg problem [37], without labelling all frames, how does one know which ones are representative? Intuitively, the diversity between the selected frames should be as large as possible. Therefore, we employ a simple but effective way; we perform a k-means clustering on the samples and randomly select equal number of frames from each class as the labelled training data. Then, we exploit the sequential unlabelled frames to create matrix D and solving Eq. (14). SSEN weights are used to predict the count on unlabelled images and the most confident unlabelled instances with predicted labels are added to the initial labelled training set. SSEN is re-trained on the updated training set and all the steps would be repeated till enough training samples being annotated.

5. Experimental setup

To evaluate the performance of the proposed method, we carry out a series of experiments on two benchmark pedestrian datasets. The UCSD dataset [6], as the largest pedestrian dataset in terms of number of frames, was collected from two viewpoints overlooking a pedestrian walkway. Example frames from both viewpoints are shown in Fig. 2. Peds1 is an oblique view including around 33,000 frames with a high people count (0–46) and Peds2 is a side-view including 34,000 frames (0–15 people). Peds1 contains crowded scenes and is substantially more challenging than Peds2 because of travelling bicycles, skateboarders and golf carts. In order to evaluate the accuracy of estimations, two widely used metrics are employed; Mean Absolute Error (MAE) and Mean Squared Error (MSE):

$$\text{MAE} = \frac{1}{N_{test}}\Sigma_{i=1}^{N_{test}}|\hat{c}_i - c_i| \quad \text{MSE} = \frac{1}{N_{test}}\Sigma_{i=1}^{N_{test}}(\hat{c}_i - c_i)^2 \tag{15}$$

where c_i and $\hat{c}_i$ are the actual and estimated counts of ith frame and N_{test} is the total number of test frames. So, to perform the evaluations, we need a reliable ground-truth as the actual counts; however, in UCSD dataset, just the first 4000 frames of each dataset have already been used for ground-truth annotation. So, we exhaustively annotated the whole dataset on the same region of interest.

In addition to UCSD dataset, we would consider another widely used dataset for people counting; the PETS 2009 dataset [38] which contains three parts showing multi-view sequences containing pedestrians walking in an outdoor environment. This dataset includes scenarios specifically created for a contest on people counting and density estimation. In the challenge, authors are required to report the number of individuals within specific region of interest for each frame for dataset S1, view 1 only. Dataset S1 has three parts; S1-L1, S1-L2 and S1-L3 and the task is to report the people count within regions (R0, R1 and R2), (R1 and R2) and R1 respectively [39]. An example frame of different scenarios is shown in Fig. 3, along with the three regions of interest. The ground-truth was generated by manually counting people in the specified regions at each sampled frame. Although PETS 2009 is not a large dataset; a wide variety of methods have been proposed and tested for this

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

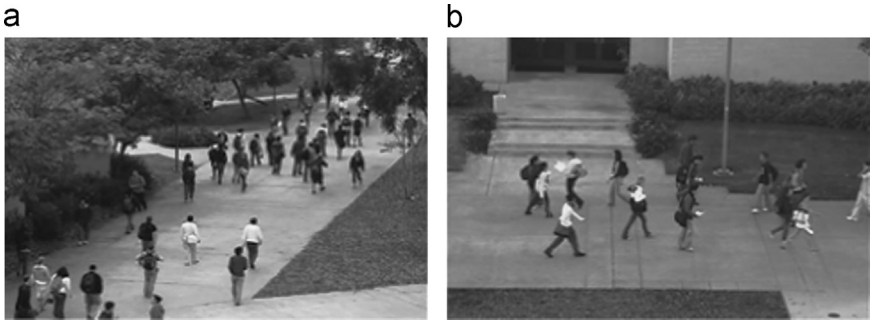

Fig. 2. UCSD pedestrian dataset. (a) Peds1 and (b) Peds2.

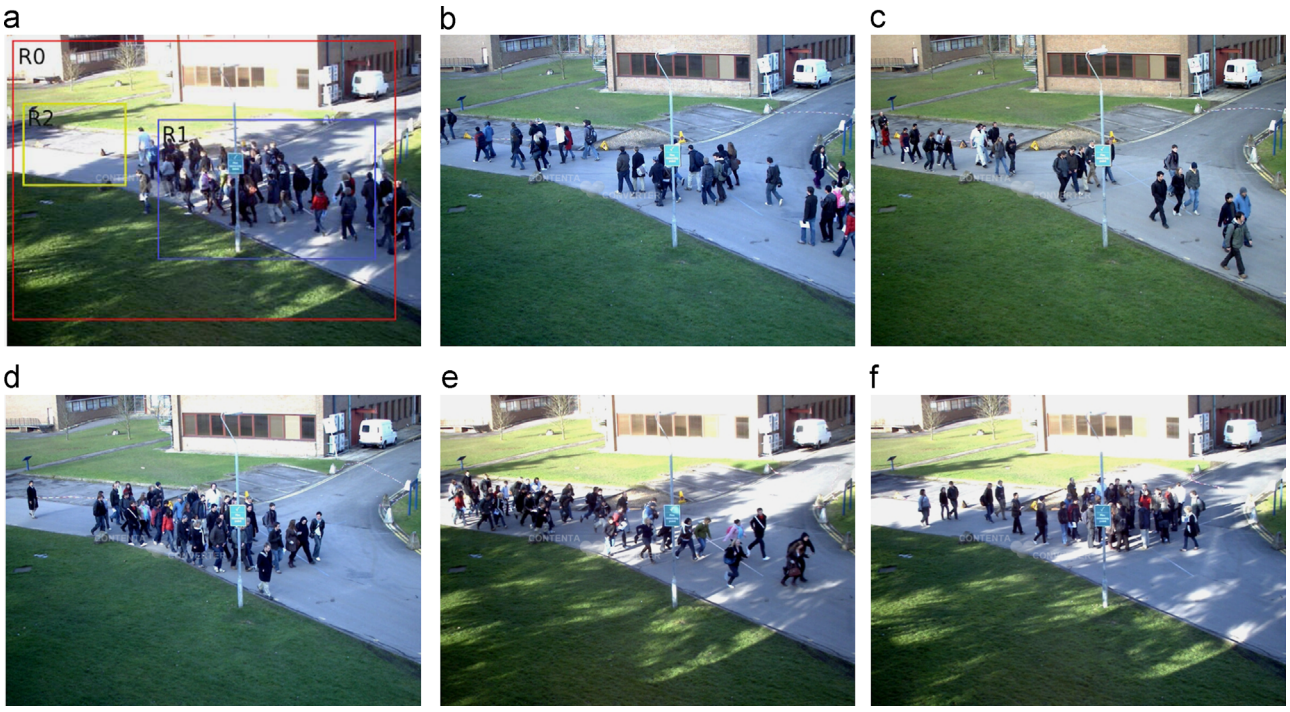

Fig. 3. PETS 2009 pedestrian dataset. (a) Regions of interest, (b) S1L1-1 Sequence, (c) S1L1-2 Sequence, (d) S1L2-1 Sequence, (e) S1L3-1 Sequence, (f) S1L3-2 Sequence.

competition and we would like to evaluate our proposed method on this dataset as well.

6. Results and discussion

6.1. Crowd counting results on UCSD dataset

We conduct extensive experiments to validate our proposed method in different scenarios. For all the experiments, training samples are randomly selected from the pedestrian dataset and the remaining ones are used for testing. Randomly choosing the training set ensures that our results and conclusions will not depend on any special choice of training data. In all the following graphs, the errors are shown in a logarithmic scale. Firstly, we examine the role of features for creating dictionary within our framework. We compare the performance on using different types of feature descriptors including Gist, HOG, CNN features obtained by DeCAF and raw sub-sampled binary image. Fig. 4 shows the MAE across various feature spaces in conjunction with five different training set sizes including 10k, 15k, 20k, 25k and 30k images on Peds1 and Peds2 datasets. According to this figure, Gist has the best overall performance over all training sizes in both datasets and it is closely followed by HOG and DeCAF. Interestingly DeCAF works well; although not better than Gist and it confirms the previous studies which claim that the activations invoked by an image within the top layers of a large CNN provide a high-level descriptor of the visual content of the image. These results suggest that instead of learning a counting model from scratch in every new scene, the labelled data from other scenes or even from a complete different dataset could be exploited to compensate for the lack of labelled data in the new scene. To put it differently, our proposed people counting method is independent of both training dataset and feature representation which means the choice of an "optimal" feature transformation is no longer critical; even CNN features trained on a different dataset, should perform as well as any other carefully hand-engineered features. Furthermore, we observe that all the descriptors gain from involving more training samples. The reason is that, when we have more diversity in the training images, the new test image can be well represented as a sparse linear combination of the training set; which in turn leads to lower estimation error. Fig. 5 displays the crowd count estimates using 30k training images and Gist feature representation on Peds1 dataset; these estimates track the ground-truth well in most of the test set. As discussed earlier, in order to reduce the complexity of solving l_1-minimization problem, random projection is employed as a fast and simple method to project the data into a much lower-dimensional space. Here, we demonstrate the

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

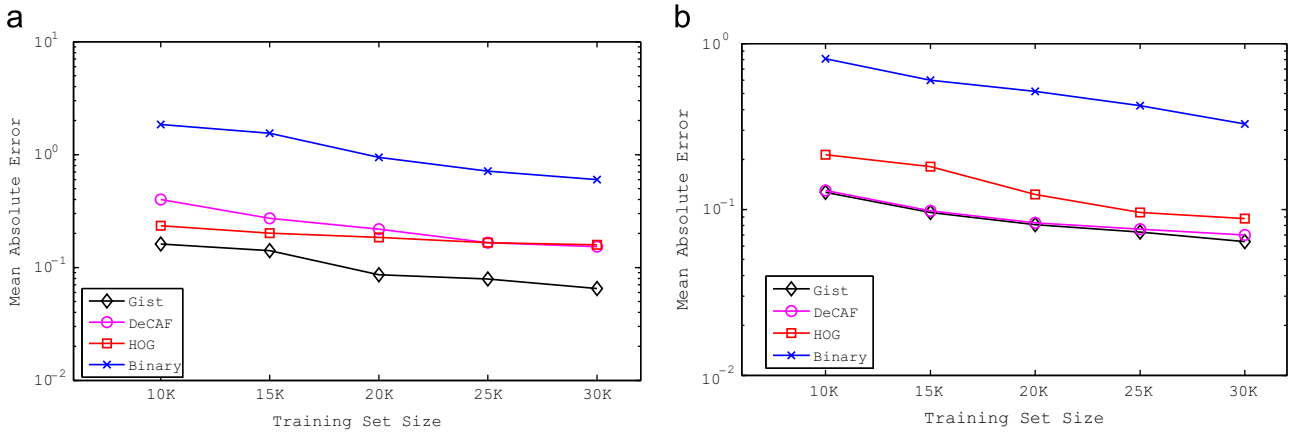

Fig. 4. MAE rate of SR-based method over various features and training sizes. (a) Peds1 and (b) Peds2.

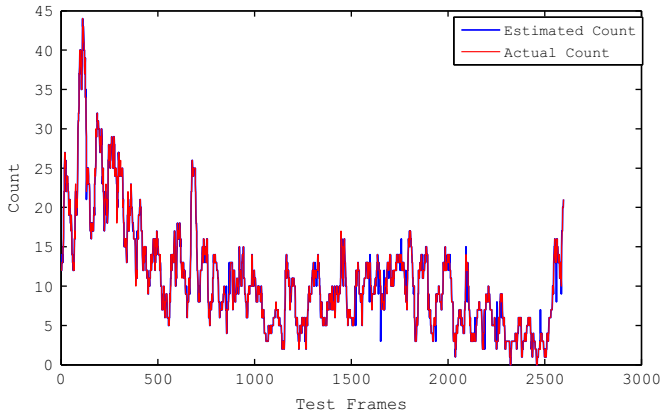

Fig. 5. Count estimation result on Peds1 dataset.

robustness of RP to preserve the similarities of the data points, while reducing the original dimension, across different feature representations in Fig. 6. We fix the training set size as 30k and use the Gaussian random matrix for the projections. We can see that Gist still performs the best in all the reduced dimensions in both datasets; also the projected vectors retain their original performance for up to even 90% reduction in the original dimension. Surprisingly, RP preserves the similarity of feature vectors well even when the data is projected to as low as 10% of original dimensions. Using such an efficient dimensionality reduction technique, the l_1-minimization can be solved much faster without losing too much accuracy. Our next experiment concerns with the evaluation of RP vs. PCA as a baseline dimensionality reduction technique, in terms of accuracy. Since Gist is the most reliable and remarkable feature to represent the crowd density in our SR-based counting method, we decided to continue the upcoming experiments with it. Here, both RP and PCA are used to construct the lower-dimensional space on different dimension fractions of Gist using 30k training images; the accuracy (MAE) of the estimations has been summarized in Fig. 7. It is observed that the error rate is less affected using RP when dimension is reduced, in comparison with PCA. This is mainly due to the fact that PCA is dependent on the omitted eigenvalues, while RP continues to give accurate results in lower dimensions. It seems that Gist feature descriptors lie on a manifold that PCA is unable to handle it gracefully. We should also note that RP has much less time complexity than PCA which in total makes it the best candidate for dimensionality reduction in our people counting task. We then study the performance of two fast l_1-solvers; Homotopy and DALM. We evaluate the efficiency of these solvers on large training sets including 10k, 15k, 20k, 25k and 30k images. In this experiment, Gist feature descriptors of selected images are used to construct the dictionary. The MAE rate of estimations can be observed in Fig. 8. Homotopy produces smaller errors and outperforms DALM in both datasets and converges in fewer steps. The objective of next experiment is to compare different methods of count estimation described in Section 3.3. It can be observed in Fig. 9that using minimum residual error in Eq. (10) slightly outperforms using maximization l_1-ratio in Eq. (9); however, these two techniques are both much better than averaging largest entries (peaks). In this experiment, Gist feature descriptors of 30k images are used to construct the dictionary.

6.2. Undersampled SRC

In all of the above experiments, we assumed that we have sufficient labelled training samples, which is achieved via manual annotation of the UCSD dataset. So as to significantly reduce the amount of manual annotation and make our method much more applicable in practice, we use SSEN which enables us to annotate very few images. The goal of the following experiments is to evaluate the effectiveness of exploiting semi-supervised learning framework and especially SSEN for automatic labelling.

We start with a small dataset and choose l labelled samples from the ground-truth (g) by performing k-means clustering on the samples; then u sequential unlabelled frames are selected and the rest of samples ($g-l$) is used as the test set. We evaluate the inductive inference performance [37] of SSEN which is the error rate on the "unlabelled data in the test partition". We examine the effect of labelled and unlabelled data by measuring the MSE performance across labelled set $\{10, 50, 100, 200, 400, 600\}$ given unlabelled set $\{0, 200, 400, 800, 1000\}$ on Peds1 and Peds2 datasets. Images are represented by their Gist feature descriptor and all the λ parameters of SSEN are chosen by a 5-fold cross validation. Fig. 10 shows clearly that when the number of labelled samples is small, increasing the number of unlabelled samples remarkably improves the annotation performance; which means that manual labelling work can be greatly reduced without losing the performance. We follow an iterative procedure aiming to obtain an enlarged labelled dataset, in which we accept that the predictions tend to be correct. During each iteration, the unlabelled samples in the test partition are given predicted labels and then the most confident unlabelled samples, together with their predicted labels, are used to enlarge the training set. Ideally, these selected unlabelled instances can finally help to learn a better classifier. The learner is re-trained on the updated training set and the whole

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

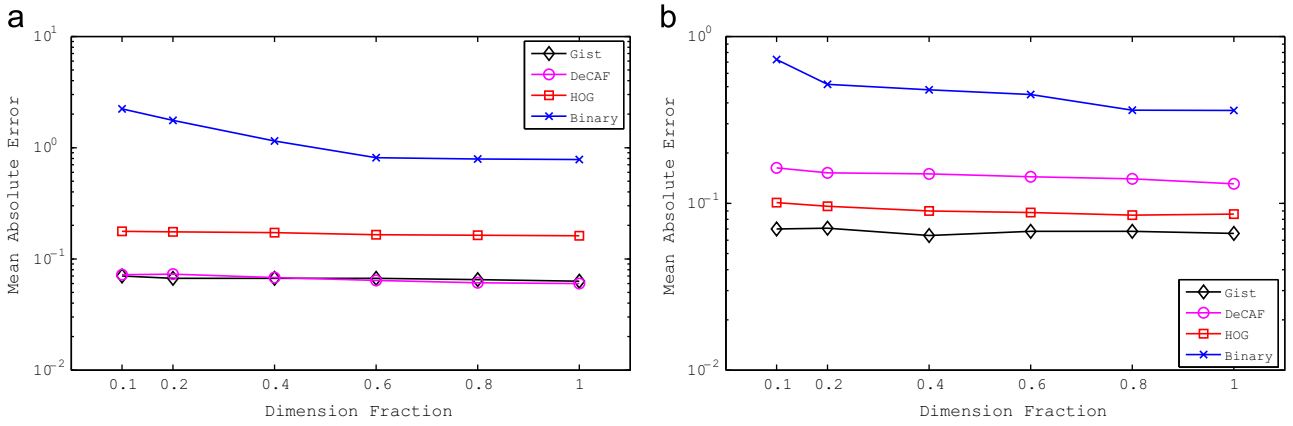

Fig. 6. MAE rate of SR-based method using RP over reduced dimensions and features. (a) Peds1 and (b) Peds2.

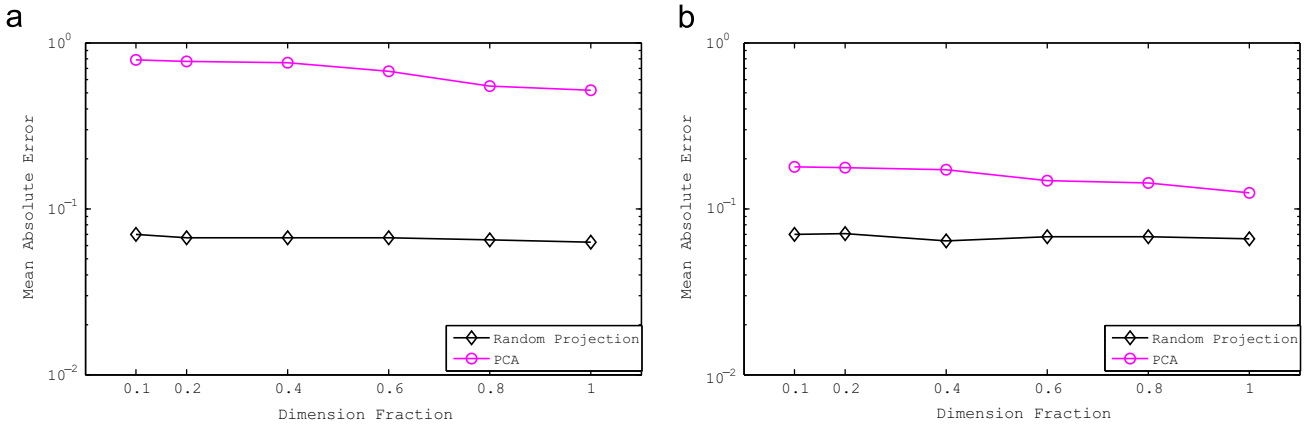

Fig. 7. MAE rate of SR-based method using RP and PCA over reduced dimensions. (a) Peds1 and (b) Peds2.

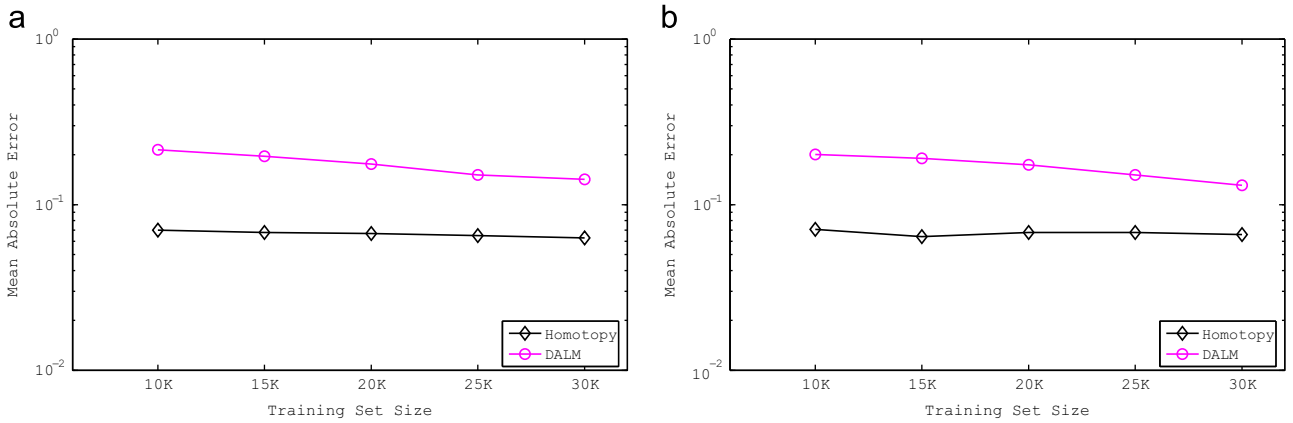

Fig. 8. MAE rate of SR-based method using Homotopy and DALM over training sets. (a) Peds1 and (b) Peds2.

process iterates until stopping criterion is met; which in our case is providing sufficient labelled samples for SRC. If SSEN wrongly assigns labels to some unlabelled instances, the final performance will be jeopardized due to the accumulation of mislabelled data. For measuring the confidence of self-labelled samples, we utilizes the information of the neighbors of each instance to identify and remove the mislabelled examples; hence a less noisy training set is obtained. Specifically in our application, the self-labelled samples are sequential frames that pedestrian quantities of each frame in a p-frame subset are the same or change slightly. So, any significant change between the predicted counts of neighboring self-labelled frames is considered as a mislabelled example. Once the possibly mislabelled examples are identified, we simply discard them and keeping the good ones intact and the filtered enlarged labelled dataset would be used for the next iteration. It is worth noting that we do not try to re-label the identified mislabelled examples essentially in the next iteration. Since most of the neighbors of a mislabelled example have been correctly labelled and moved to enlarge the labelled training set; the sudden prediction change could not be exploited any more due to the lack of neighbors, in the next iteration. Alternatively, mislabelled examples would be held out and annotated in the final iteration, in which our model

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

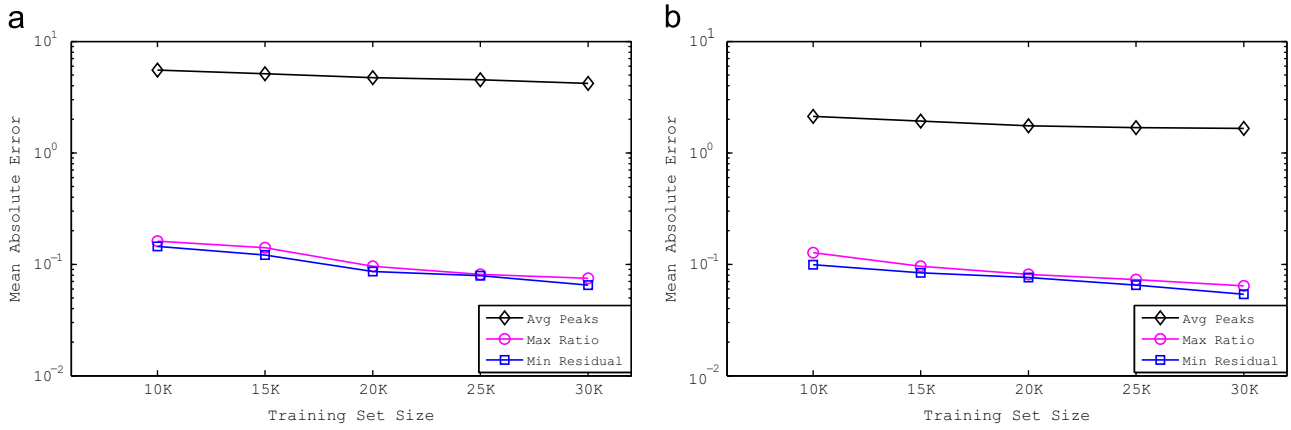

Fig. 9. MAE rate of SR-based method using different count estimation techniques. (a) Peds1 and (b) Peds2.

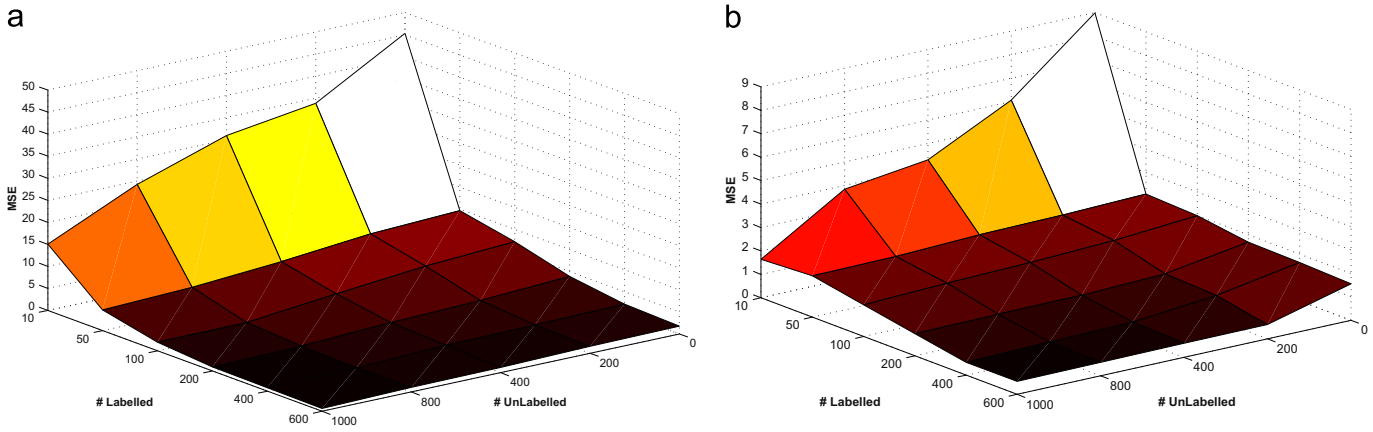

Fig. 10. The effect of labelled and unlabelled data in SSEN. (a) Peds1 and (b) Peds2.

has been trained sufficiently with more confident labels of their neighboring frames and would assign correct label to them with high probability. In each iteration, 40% data are kept aside to evaluate the performance of learned hypothesis, while the remaining 60% data are partitioned into labelled set and unlabelled set under the unlabel rate 75%, i.e. just 25% (of the 60%) data are used as labelled examples while the remaining 75% (of the 60%) data are used as unlabelled examples.

This iterative self-training labelling procedure is very beneficial when the scale of initial labelled training samples is too small to train SSEN with good generalization. Mislabelling of such large amount of unlabelled data is unavoidable and SSEN or even any other regression model would generate poor results. Since we wish to keep the manual labelling effort minimum without sacrificing accuracy, we start with a small labelled dataset, and as the unlabelled training data sequentially arrives, we use SSEN to predict the counts on them. The confident self-labelled samples are added to the initial set and we treat them as labelled data; consequently lack of enough labelled samples becomes less a problem. One question maybe raised here; if iterative SSEN is a reliable method for predicting frame count, why not use it to annotate (and meanwhile estimate the counts) the whole dataset rather than utilizing SSEN just for preparing the rich training set of SRC? Although SSEN provides promising results on small datasets, it is more prone to erroneous prediction on large-scale datasets in comparison with SRC. Regression models, including SSEN, suffer from serious problems such as poor tractability and expensive training time, when they are generalized to large-scale datasets. The iterative SSEN also needs couple of training iteration which itself imposes more error and computational complexity. Importantly, SSEN relies on the assumptions that the temporal space is dense and abundant sequential unlabelled frames are available; however this assumptions can be too stringent for many real-world scenarios when continuous video recording is not available.

We would like to examine the accuracy of the self-labelled samples, so we repeat the SRC experiments this time with predicted counts obtained by SSEN instead of the manual ground-truth (actual counts). Tables 1 and 2 summarize the MAE of supervised and semi-supervised SRC across five different training set sizes including 10k, 15k, 20k, 25k and 30k on Peds1 and Peds2 datasets while using Gist and DeCAF feature representations. To have a fair comparison, the training and test partitions of supervised and semi-supervised methods are kept similar in all the experiments. Note, to achieve 30k labelled images of semi-supervised SRC, just a total of 600 labelled images have been initially used and then by exploiting iterative SSEN, the labelled set would be enlarged gradually. Not surprisingly, the counting accuracy is higher in the supervised SRC and this can be explained by mislabelled examples and error accumulation; however, the performance has not been affected significantly, confirming the robustness of sparse representation to the noise introduced in the self-labelling process. Gist still has the best overall performance in the semi-supervised SRC over all training sizes in both datasets.

6.3. Comparison with other methods

To demonstrate the effectiveness of SR-based people counting method on large-scale datasets, extensive comparative evaluations are conducted against several popular methods. The goal of these

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

Table 1
Comparison of supervised and semi-supervised SRC on Peds1 dataset.

Method	Feature	MAE					MSE				
		10k	15k	20k	25k	30k	10k	15k	20k	25k	30k
Sup.	Gist	0.16	0.14	0.08	0.07	0.06	0.23	0.13	0.09	0.08	0.06
Semi-Sup.	Gist	0.25	0.20	0.17	0.15	0.13	0.35	0.30	0.24	0.19	0.16
Sup.	DeCAF	0.39	0.27	0.21	0.16	0.15	1.09	0.65	0.43	0.30	0.22
Semi-Sup.	DeCAF	0.50	0.48	0.45	0.38	0.31	1.10	0.99	0.87	0.80	0.75

Table 2
Comparison of supervised and semi-supervised SRC on Peds2 dataset.

Method	Feature	MAE					MSE				
		10k	15k	20k	25k	30k	10k	15k	20k	25k	30k
Sup.	Gist	0.12	0.09	0.08	0.06	0.07	0.24	0.12	0.10	0.07	0.10
Semi-Sup.	Gist	0.25	0.21	0.17	0.14	0.10	0.31	0.26	0.23	0.20	0.16
Sup.	DeCAF	0.12	0.09	0.07	0.07	0.07	0.14	0.10	0.08	0.07	0.07
Semi-Sup.	DeCAF	0.24	0.21	0.17	0.15	0.11	0.29	0.22	0.18	0.14	0.10

Table 3
Counting accuracy of different methods over various training sizes on Peds1 dataset.

Method	Feature	MAE					MSE				
		10k	15k	20k	25k	30k	10k	15k	20k	25k	30k
SR-RP	Gist	0.16	0.14	0.08	0.07	0.06	0.23	0.13	0.09	0.08	0.06
SR-RP	DeCAF	0.39	0.27	0.21	0.16	0.15	1.09	0.65	0.43	0.30	0.22
Regression	SET	8.72	6.76	5.89	5.69	5.42	97.84	66.23	56.02	53.93	54.33
SR-Pooling	SIFT	9.73	8.39	7.59	7.37	7.10	149.10	123.33	105.01	99.70	90.61
NN	Gist	2.90	2.26	2.17	1.76	1.62	13.28	9.12	8.50	5.64	5.09
Lin-SVM	Gist	1.03	1.01	1.01	1.02	1.01	2.14	2.11	2.08	1.99	2.03
RBF-SVM	Gist	0.23	0.15	0.12	0.11	0.10	0.31	0.25	0.21	0.19	0.17
BOF	SIFT	9.75	8.31	7.74	7.42	7.20	152.53	119.24	107.31	100.30	95.21
VLAD	SIFT	9.73	8.38	7.52	7.30	7.14	150.41	120.74	104.26	100.14	95.37
Fisher	SIFT	9.85	8.40	7.57	7.35	7.01	152.10	123.41	104.96	99.27	90.54

Table 4
Counting accuracy of different methods over various training sizes on Peds2 dataset.

Method	Feature	MAE					MSE				
		10k	15k	20k	25k	30k	10k	15k	20k	25k	30k
SR-RP	Gist	0.12	0.09	0.08	0.06	0.07	0.24	0.12	0.10	0.07	0.10
SR-RP	DeCAF	0.12	0.09	0.07	0.07	0.07	0.14	0.10	0.08	0.07	0.07
Regression	SET	2.48	2.05	1.91	1.88	1.82	8.26	6.23	5.57	5.55	5.54
SR-Pooling	SIFT	3.10	2.62	2.54	2.39	2.21	16.03	12.55	11.93	10.70	9.34
NN	Gist	0.40	0.37	0.31	0.17	0.10	0.64	0.60	0.50	0.23	0.15
Lin-SVM	Gist	0.30	0.29	0.29	0.30	0.26	0.34	0.33	0.33	0.33	0.32
RBF-SVM	Gist	0.15	0.13	0.12	0.11	0.10	0.24	0.18	0.16	0.15	0.13
BOF	SIFT	3.30	2.70	2.49	2.42	2.10	15.60	12.91	11.34	10.92	9.93
VLAD	SIFT	3.04	2.60	2.55	2.39	2.10	15.61	12.41	12.09	10.49	9.23
Fisher	SIFT	2.97	2.68	2.45	2.37	2.11	15.00	13.18	11.30	10.56	9.35

experiments is to evaluate the capability of the available methods to deal with large-scale datasets and verify their generalization power. In all the upcoming experiments, training samples are randomly selected from the datasets, then selected feature representation is used to construct the dictionary and the rest of images are used for testing purpose. For our proposed method, RP is then employed and just 10% of original feature vector dimension is used. Tables 3 and 4 compare the accuracy (MAE/MSE) of all the following methods for estimating people count using different training set sizes on Peds1 and Peds2 datasets respectively.

First, we compare our approach with a state-of-the-art counting by the regression method. We select features and the regression method that are both representative and promising in terms of originally reported performance [6]. The combination of three type of low-level global features has been widely used in the related papers; Segment (e.g. area, perimeter), Edge (e.g. internal edge length and orientation) and Texture (e.g. Entropy) features; which we call them SET here. These features are extracted from crowd segments, perspective normalised [6] and a feature vector is formed by concatenating them which is used as the input for the regression

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

model. We select Gaussian Process Regression (GPR) which was reported as one of the most accurate ones to handle the crowd density [6]. Although regression models could achieve promising results on small datasets, they suffer from serious weaknesses when they are generalized on large-scale datasets. Loy et al. [8] reveal that the actual performance of regression models can be quite different from what we anticipate, especially when they are applied to an unseen density, they tend to overestimate or underestimate subject to nature of data. Another key weakness is their poor tractability to large training datasets. Usually when more and more data is involved they are unable to adequately capture the non-linear trend in the feature space. They also need online training which takes a lot of time, in particular the time complexity of GPR is bounded by $O(n^3)$ which is a limitation in large datasets. Besides, selecting the optimal feature combination is an important step which is very dependant on the crowd structure and density.

We observed that gradient-based global descriptors (especially Gist) perform well in predicting count in large-scale datasets; this motivated us to examine the effectiveness of local gradient-based descriptors in the sparse representation framework as well. Initially, some local descriptors like SIFT [15] are extracted from an image and then each local descriptor is encoded into a sparse code using a dictionary learning technique. After obtaining the sparse codes, we pool them into a single vector using max pooling. Ge et al. [40] used this idea to derive a compact yet discriminative image representation from multiple types of features for large-scale image retrieval and achieved significant results on benchmark retrieval datasets; however, the aggregated local features cannot capture the crowd patterns properly.

We next compare SRC with three classical classifiers, namely, NN, linear SVM and non-linear SVM with RBF kernel within the context of people counting on large-scale datasets. We can observe that SRC outperforms NN in both datasets and this is essentially because NN independently evaluates the distance between the test sample and one training sample; in contrast, SRC uses a linear combination of all the training samples to represent the test sample and classify it into the class with the minimum deviation and this justifies SRC achievement. Essentially, the training samples are not uncorrelated and the distance between the test and training samples should not be independently calculated, rather; the relationship between different training instances should be taken into account. NN performs a linear search and adopts only one sample in $O(n)$; while SRC finds the sparsest representation in $O(n^2)$. The multi-class SVM classifier is implemented using the LIBSVM [41], which uses a one-against-one decomposition strategy. We select two kernels, linear and Gaussian radial basis function (RBF). Linear SVM is not appropriate for separating Gist features of different crowd densities; nevertheless, better performance is achieved using non-linear kernels like RBF; however our method is still better than best of SVM. When the classes in training data are non-separable by the SVM, SRC might have an advantage over the SVM classifier, depending on similarity between the test vector and the training exemplars from the same class. All the parameters and also the best model are estimated through 5-fold cross-validation over a large training set and this step is really time-consuming. Also, the time complexity of non-linear SVM is generally between $O(n^2)$ and $O(n^3)$ depending on the number of iterations [41]; however this depends a lot on the solving techniques. In brief, SRC outperforms traditional classifiers, meanwhile, it is faster and does not need any model/parameter selection. Finally, the proposed SR-based method is compared with state-of-the-art image retrieval methods including Bag of Features (BOF) [42] and aggregation-based representation, i.e. Fisher Vector [43] and VLAD [44]. In the former, a set of local image patches is sampled using a keypoint detector and a vector of visual descriptors is evaluated on each patch independently. The resulting distribution of descriptors is then quantified to convert it to a histogram of votes for codebook centers and the resulting global descriptor vector is used as a characterization of the image. In Fisher and VLAD methods, a bag of local features in an image is converted into a global, fixed size vector representation through vector quantization. We observe that the evaluated retrieval methods, easily fail to estimate crowd density in all large-scale datasets and global image descriptors outperform the local ones in general. We also compare the average processing time spent on a test image to estimate its count using different methods on Peds1 dataset while using 30k training images in Table 5; notice that the training time has not been considered here.

On the whole, results suggest that SR-based people counting method is superior to all the other evaluated techniques and the errors of our method are significantly less than the others' including state-of-the-art counting by regression. In brief, the best performance can be achieved by using the Gist descriptor for creating the dictionary, Homotopy l_1-minimizer and minimum-residual error as the count estimation method; meanwhile random projection is exploited as the dimensionality reduction technique.

6.4. SRC on small datasets

The proposed SR-based people counting method performs well in predicting counts in large datasets; however, we are also interested in evaluating it on smaller datasets, which are popular in counting by regression papers. We follow the experimental protocol in [6], where the training set consists of 1200 and 1000 frames for Peds1 and Peds2 datasets respectively, and the remaining 2800 and 3000 frames are held out as the test set. Table 6 presents the counting accuracy of our method versus [6] as the state-of-the-art regression-based method. Although the error rates are not as small as the larger datasets (e.g. 30k images); our proposed method still outperforms the regression-base method, even with CNN feature representation. Obviously, as the data keeps getting bigger, SRC is coming to play a key role in providing better estimations and the superiority of SRC would be more remarkable. Notice that these conditions are different from the "Undersampled SRC"; here the dictionary is rich enough to span variations of different classes under testing conditions because the

Table 5
Average processing time of a test image in seconds across different methods.

Method	Time	Details
SR-RP	0.55	Gist extraction: 0.1, Test time (l_1-min): 0.45
Regression	2.10	SET extraction: 1, GPR test time: 1.10
SR-Pooling	0.80	Encoding: 0.6, SIFT extraction: 0.2, Query time: 0.005
NN	0.60	Gist extraction: 0.1, Linear search: 0.5
Lin-SVM	0.30	Gist extraction: 0.1, SMV test time: 0.2
RBF-SVM	0.30	Gist extraction: 0.1, SVM test time: 0.2
BOF	1.84	Encoding: 1.55, SIFT extraction: 0.2, Query time: 0.09
VLAD	0.37	Encoding: 0.15, SIFT extraction: 0.2, Query time: 0.02
Fisher	0.86	Encoding: 0.16, SIFT extraction: 0.2, Query time: 0.5

Table 6
Counting accuracy of SR-based and regression-based methods on UCSD dataset.

Method	Feature	Peds1		Peds2	
		MAE	MSE	MAE	MSE
SR-RP	Gist	0.55	4.13	0.74	2.03
SR-RP	DeCAF	0.87	6.69	0.97	3.62
GPR [6]	SET	3.65	7.41	1.58	2.16

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

test set is not very big and crowd patterns of training/test sets are not too different. So, there is no need to enlarge the labelled training set any more.

6.5. Crowd counting results on PETS 2009 dataset

We conduct some experiments to validate our proposed method on PETS dataset as well. We would compare the performance of different feature descriptors, including Gist, HOG and CNN features obtained by DeCAF to represent the crowd density in the proposed SR-based counting method. According to the best results obtained on UCSD dataset, we would exploit Homotopy method as the l_1-minimizer and RP is also employed for dimensionality reduction and just 10% of original feature vector is used. First, we compare the proposed SR-based counting method with all the methods described earlier, i.e. regression method, retrieval methods, general classifiers. Table 7 reports the counting accuracy (MAE) of count estimations on PETS dataset. For each test video sequence, the task is to report the count on specified regions of interest. Training images are selected from other sequences available in the PETS dataset that were not used for the test purpose.

Similar to the results obtained on UCSD dataset, we observe that our method outperforms all the other evaluated methods. In this dataset, Gist still has the best overall performance through all sequences and it is closely followed by HOG and DeCAF. CNN features trained on ImageNet dataset perform as well as handcrafted features; however local descriptors fail to represent the crowd density using either retrieval or sparse coding methods. SRC also outperforms traditional classifiers and regression method (GPR). Fig. 11 shows the estimated number of people using Gist feature representations with respect to time for considered regions over S1L1-1 sequences of PETS dataset. We notice that our method is able to provide a good estimate in almost all cases.

Next, we compare our method with the most successful methods of people counting participated in PETS 2009 competition; these methods have been chosen according to exhaustive performance evaluation performed in [52]. Here we briefly introduce these methods; greater details of each method can be found in the corresponding papers. For people counting, Chan et al. [45] segment the video into crowd regions moving in different directions, using a mixture of dynamic textures. Different low-level global features are extracted from each segment and the number of people per segment is estimated using Gaussian process regression. Alahi et al. [46] create degraded foreground silhouettes from some binary images and these silhouettes are then used as the atoms of a multi-silhouette dictionary to model the presence of individuals at given locations on an occupancy grid. An occupancy vector is constructed from the observed data using a re-weighted Lasso method. Albiol et al.'s [47] method is based on detecting salient points or corners, along with their motion vectors. A motion vector with respect to the previous frame is estimated using a multi-resolution block matching technique for each detected corner. Statistical measures on this data would be used to estimate the number of people. Choudri et al.'s [48] method is based on scale-weighted pixel counting to approximate the number of people in a region of interest. The foreground pixels are explicitly classified and pixels that are not of interest are updated as background and finally a head detector is used for human detection and counting. Paetzold et al. [49] present a counting by a detection method based on fusing spatial information of an adapted HOG with temporal information by exploiting distinctive motion characteristics of different human body parts. An algorithm is then applied to validate the trajectories associated with a human and a count is made of the resulting detections to identify the number of individuals in a scene. In Conte et al.'s [50] method, SURF [53] features are extracted from images, which is followed by a Support Vector Regression (ϵ-SVR) function to estimate the number of people present in the scene. Subburaman et al. [51] employ a head detector based on boosted cascade of integral features. To prune the search region for the head detector, an interest point detector based on gradient orientation feature is applied to locate regions similar to the top of head region from gray level images.

Table 8 compares the MAE of our proposed method against aforementioned counting methods. As it can be inferred, the

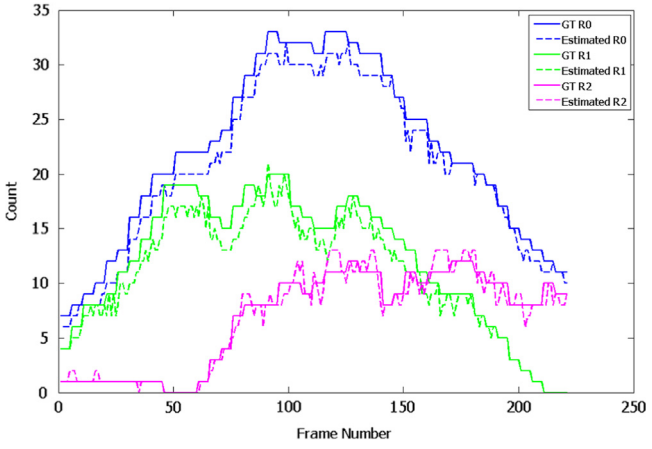

Fig. 11. Count estimation result on PETS dataset (S1L1-1).

Table 7
Counting estimation error of different methods on PETS 2009 dataset.

Method	Feature	Sequence									
		S1L1-1			S1L1-2			S1L2-1		S1L3-1	S1L3-2
		R0	R1	R2	R0	R1	R2	R1	R2	R1	R1
SR-RP	Gist	1.30	1.23	0.71	1.01	0.70	0.99	1.95	1.90	0.86	0.01
SR-RP	HOG	1.80	1.87	0.91	1.79	0.96	1.31	2.32	2.43	1.01	0.41
SR-RP	DeCAF	2.10	2.14	1.10	2.08	0.96	1.30	2.50	2.63	1.20	0.86
Regression	SET	6.35	5.43	3.15	4.11	2.76	1.34	9.73	7.65	7.19	4.77
SR-Pooling	SIFT	9.23	5.21	4.02	6.93	3.60	4.04	9.65	6.25	7.87	3.97
NN	Gist	3.60	2.13	1.10	5.11	2.10	1.01	6.51	2.50	4.76	0.10
Lin-SVM	Gist	3.41	2.00	1.09	5.13	1.90	1.29	6.12	2.64	4.64	1.10
RBF-SVM	Gist	3.44	2.10	0.99	5.05	1.90	1.05	6.38	2.73	4.75	0.93
BOF	SIFT	10.01	8.21	3.76	7.24	2.10	4.87	12.67	5.89	7.10	3.76
VLAD	SIFT	8.45	7.11	3.42	7.13	2.00	5.10	9.59	5.10	7.68	2.89
Fisher	SIFT	10.76	8.43	4.99	7.97	2.54	5.03	13.34	9.76	8.55	4.08

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

Table 8
Crowd counting estimation error of different counting methods on PETS 2009.

Sequence	Region	Method							
		SR-RP (Our)	Chan [45]	Alahi [46]	Albiol [47]	Choduri [48]	Patzold [49]	Conte [50]	Subburaman [51]
S1L1-1	R0	1.30	2.46	–	1.42	1.29	2.75	1.38	5.95
	R1	1.23	2.28	–	–	2.23	2.58	2.14	1.90
	R2	0.71	0.99	–	–	0.70	1.38	7.60	2.50
S1L1-2	R0	1.01	1.41	4.20	1.77	3.26	2.35	1.14	2.08
	R1	0.70	0.69	2.30	–	3.18	1.58	0.80	1.86
	R2	0.99	1.23	1.87	–	1.04	1.58	0.87	0.86
S1L2-1	R1	1.95	5.89	6.50	1.94	3.70	6.37	2.18	2.40
	R2	1.90	4.48	4.00	–	4.17	6.08	3.25	1.40
S1L3-1	R1	0.86	0.98	0.90	1.36	0.67	4.70	2.95	7.00
S1L3-2	R1	0.01	–	8.83	–	16.50	24.38	–	16.00

proposed SR-RP method performs robustly throughout all the sequences and especially shows a promising performance on very dense crowd sequences such as S1L2-1 and S1L3-1.

6.6. Discussion and future work

Discriminative feature representation plays an important role for achieving the state-of-the-art performance in image classification. In particular, learning spare representation has recently achieved impressive results and found many applications in machine learning and computer vision fields. Sparse coding represents an input data as a linear combination of a few items from a dictionary. The performance of sparse representation is closely related to the dictionary, which should faithfully and discriminatively represent the test image. The SRC algorithm naively uses all the training samples as the dictionary and achieves promising performance in many applications including people counting. Improvements in computational and memory efficiency of the SRC method make it a more practical solution to be implemented for real-time applications. To this aim, in this paper, we utilized fast l_1-solvers and dimensionality reduction technique. For some applications, however, rather than using the entire set of training data as dictionary, it is computationally advantageous to learn a compact dictionary from training data. Different algorithms have been developed for learning such dictionaries in an unsupervised and supervised fashion; however recent research indicates that the dictionaries obtained in an unsupervised way, may not necessarily be the best for classification [54].

Many efforts have been dedicated to embedding the discriminative information into the representations via supervised learning [54]. Supervised learning approaches can be divided into different categories [55]: while the algorithms in the first group, learn multiple or class-specific dictionaries to promote discrimination, the second set of approaches incorporate discriminative terms into the objective function of dictionary learning and the last type of approaches learn a compact dictionary by merging or selecting dictionary items from an initially large dictionary. Taking a step further, some researchers designed dictionaries for the situations that the present training instances are different from the testing instances. For instance, Zhu et al. [56] utilized weakly labelled data from other visual domains as the auxiliary source data for enhancing the original learning system and their framework only requires a small set of labelled samples in the source domain.

Nevertheless, most dictionary learning algorithms are of high time complexity and converge slowly; moreover, they may get trapped in local minimum [55]; consequently, efficient dictionary learning on large-scale data still remains a challenging task. It is worth noting that submodular optimization has become a sensible trend to solve large-scale problems in computer vision and as an example, Jiang et al. [55] exploited submodularity and monotonicity properties of object function to construct a dictionary from a set of dictionary item candidates. Possible future work includes exploring supervised dictionary learning methods to propose an efficient method to learn a compact and discriminative dictionary for large-scale people counting application. When the dictionary is large and the data dimension is high, learning sparse representations is a computationally challenging problem and this is the direction in which we would like to extend our work.

Furthermore, although the literature of image classification is predominated by local and global hand-crafted features; deep learning methodologies have been utilized recently to obtain machine-learned features for image classification and have shown great potential in various visual recognition tasks. In this paper, the features obtained by CNN yields encouraging results. Meanwhile, various architectures and techniques have been proposed to enhance the learning capacity. Recently, Shao et al. [57] proposed a multispectral neural network to learn features from multicolumn deep neural networks to obtain an effective low-dimensional embedding, which led to a more discriminative feature than that of CNN. However, deep architectures often require a large amount of labelled data for supervised training; their training would take very long time and a large number of hyper-parameters should be tuned. Alternatively, proposing an optimal solution can be considered as a generalized way to extract the most meaningful features for any user-defined application. As a successful example, in [58] the authors developed an evolutionary learning methodology to automatically generate domain-adaptive global feature descriptors for image classification using multiobjective genetic programming. An intriguing question for future work is whether this framework could be useful for people counting/detection applications.

7. Conclusion

In this paper, we proposed an extremely accurate and scalable people counting method based on sparse representation. Sparsity provides a powerful tool for inferring high-dimensional image data that have complex low-dimensional structure. Methods like l_1-minimization offer computational tools to extract such structures and help harness the semantic of data. In order to reduce the computation complexity of l_1-minimization solvers for finding such sparse representation, random projection is employed as a fast and simple dimensionality reduction method which preserves the similarities of the data vectors well. According to our extensive

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

experiments on two benchmark datasets, the proposed SR-based people counting method achieves the best performance in comparison with all evaluated methods including counting by regression techniques. The key to this success is choosing the dictionary in such a way that sufficient diversity is provided in training samples and exploiting the discriminative power of image descriptors especially Gist in characterizing the crowd density; however, our numerical results consistently demonstrate that if sparsity is properly harnessed, the choice of global features is no longer critical. We showed that the features extracted from a deep Convolutional network trained in a fully supervised fashion on a large, fixed set of object recognition task can be reused to other tasks such as people counting.

In addition, in order to provide the labelled training set with sufficient diversity, a semi-supervised learning framework is employed to enable image annotation with just a few labelled sample images through exploiting the sequential information of readily available vast quantity of unlabelled data. “Big data” offers great potential to exploit the discriminative power of SRC; moreover, noisy and mislabelled training set can be handled uniformly and robustly within the same classification framework.

Conflict of interest

No conflict on interest.

Acknowledgements

Authors are grateful for the research support obtained from NSERC and AQL Management Consultant Inc.

References

[1] M. Li, Z. Zhang, K. Huang, T. Tan, Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection, in: 19th International Conference on Pattern Recognition, ICPR 2008, IEEE, 2008, pp. 1–4.
[2] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: an evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 743–761.
[3] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, 2005, pp. 886–893.
[4] G.J. Brostow, R. Cipolla, Unsupervised Bayesian detection of independent motion in crowds, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2006, pp. 594–601.
[5] V. Rabaud, S. Belongie, Counting crowded moving objects, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2006, pp. 705–711.
[6] A.B. Chan, N. Vasconcelos, Counting people with low-level features and Bayesian regression, IEEE Trans. Image Process. 21 (4) (2012) 2160–2177.
[7] C.C.L.S.G. Chen, Ke, T. Xiang, Feature mining for localised crowd counting, in: BMVC, vol. 1, 2012, pp. 3–14.
[8] C.C. Loy, K. Chen, S. Gong, T. Xiang, Crowd counting and profiling: methodology and evaluation, in: Modeling, Simulation and Visual Analysis of Crowds, Springer, New York, 2013, pp. 347–382.
[9] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, S. Yan, Sparse representation for computer vision and pattern recognition, Proc. IEEE 98 (6) (2010) 1031–1044.
[10] E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inf. Theory 52 (2) (2006) 489–509.
[11] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[12] V. Cevher, A. Sankaranarayanan, M.F. Duarte, D. Reddy, R.G. Baraniuk, R. Chellappa, Compressive sensing for background subtraction, 2008.
[13] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Discriminative learned dictionaries for local image analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, IEEE, 2008, pp. 1–8.
[14] Y.C. Eldar, G. Kutyniok, Compressed Sensing: Theory and Applications, Cambridge University Press, 2012.
[15] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[16] H. Foroughi, N. Ray, H. Zhang, People counting with image retrieval using compressed sensing, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP'14), vol. 14, 2014.
[17] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001) 145–175.
[18] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, 2009, pp. 248–255.
[20] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: a deep convolutional activation feature for generic visual recognition, arXiv preprint arXiv:1310.1531.
[21] A.S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: an astounding baseline for recognition, arXiv preprint arXiv:1403.6382.
[22] L. Maaten, E. Postma, J. Herik, Dimensionality reduction: a comparative review, J. Mach. Learn. Res. 10 (2009) 66–71.
[23] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognit. Neurosci. 3 (1) (1991) 71–86.
[24] E. Bingham, H. Mannila, Random projection in dimensionality reduction: applications to image and text data, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, CA, USA, 2001, pp. 245–250.
[25] W. Johnson, J. Lindenstrauss, Extensions of Lipschitz mappings into Hilbert space, in: Proceedings of the Modern Analysis and Probability Conference, vol. 26, AMS, New Haven, Conn , 1982, pp. 189–206.
[26] E. Candes, J. Romberg, l1-magic: recovery of sparse signals via convex programming URL: ⟨www.acm.caltech.edu/l1magic/downloads/l1magic.pdf⟩ 4, 2005.
[27] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, An interior-point method for large-scale l1-regularized least squares, IEEE J. Sel. Top. Signal Process. 1 (4) (2007) 606–617.
[28] A. Yang, A. Ganesh, S. Sastry, Y. Ma, Fast L1-minimization algorithms and an application in robust face recognition: a review, Technical Report, EECS Department, University of California, Berkeley, 2010.
[29] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Stat. 32 (2) (2004) 407–499.
[30] D.M. Malioutov, M. Cetin, A.S. Willsky, Homotopy continuation for sparse signal representation, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP'05), vol. 5, 2005, v–733.
[31] D.L. Donoho, Y. Tsaig, Fast solution of l1-norm minimization problems when the solution may be sparse, IEEE Trans. Inf. Theory 54 (11) (2008) 4789–4812.
[32] J. Yang, Y. Zhang, Alternating direction algorithms for l1-problems in compressive sensing, SIAM J. Sci. Comput. 33 (1) (2011) 250–278.
[33] S. Huanga, D. Yanga, J. Zhoua, L. Huangfub, Image classification by sparse representation-based transduction, arXiv preprint arXiv:1408.6257.
[34] W. Deng, J. Hu, J. Guo, Extended SRC: undersampled face recognition via intraclass variant dictionary, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1864–1870.
[35] B. Tan, J. Zhang, L. Wang, Semi-supervised elastic net for pedestrian counting, Pattern Recognit. 44 (10) (2011) 2297–2304.
[36] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc.: Ser. B Stat. Methodol. 67 (2) (2005) 301–320.
[37] C.C. Loy, S. Gong, T. Xiang, From semi-supervised to transfer counting of crowds, in: IEEE International Conference on Computer Vision (ICCV), 2013, 2013, pp. 2256–2263.
[38] PETS 2009, ⟨http://www.cvg.rdg.ac.uk/PETS2009/⟩, 2009.
[39] J. Ferryman, A. Ellis, PETS2010: dataset and challenge, in: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2010, pp. 143–150.
[40] T. Ge, Q. Ke, J. Sun, Sparse-coded features for image retrieval, BMVC, 2013.
[41] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27.
[42] J. Sivic, A. Zisserman, Video Google: a text retrieval approach to object matching in videos, in: IEEE International Conference on Computer Vision, 2003, Ninth Proceedings, 2003, pp. 1470–1477.
[43] F. Perronnin, Y. Liu, J. Sánchez, H. Poirier, Large-scale image retrieval with compressed fisher vectors, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, IEEE, 2010, pp. 3384–3391.
[44] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, 2010, pp. 3304–3311.
[45] A.B. Chan, M. Morrow, N. Vasconcelos, Analysis of crowded scenes using holistic properties, in: Performance Evaluation of Tracking and Surveillance workshop at CVPR, 2009, pp. 101–108.
[46] A. Alahi, L. Jacques, Y. Boursier, P. Vandergheynst, Sparsity-driven people localization algorithm: Evaluation in crowded scenes environments, in: 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter), IEEE, 2009, pp. 1–8.
[47] A. Albiol, M.J. Silla, A. Albiol, J.M. Mossi, Video analysis using corner motion statistics, in: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2009, pp. 31–38.

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009

[48] S. Choudri, J.M. Ferryman, A. Badii, Robust background model for pixel based people counting using a single uncalibrated camera, in: 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter), IEEE, 2009, pp. 1–8.
[49] M. Patzold, R.H. Evangelio, T. Sikora, Counting people in crowded environments by fusion of shape and motion information, in: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2010, pp. 157–164.
[50] D. Conte, P. Foggia, G. Percannella, F. Tufano, M. Vento, A method for counting people in crowded scenes, in: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2010, pp. 225–232.
[51] V.B. Subburaman, A. Descamps, C. Carincotte, Counting people in the crowd using a generic head detector, in: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS), IEEE, 2012, pp. 470–475.
[52] J. Ferryman, A.-L. Ellis, Performance evaluation of crowd image analysis using the PETS2009 dataset, Pattern Recognit. Lett. 44 (2014) 3–15.
[53] H. Bay, T. Tuytelaars, L. Van Gool, Surf: speeded up robust features, in: Computer Vision–ECCV 2006, Springer, Berlin, Heidelberg, 2006, pp. 404–417.
[54] J. Tang, L. Shao, X. Li, Efficient dictionary learning for visual categorization, Comput. Vis. Image Underst. 124 (2014) 91–98.
[55] Z. Jiang, G. Zhang, L.S. Davis, Submodular dictionary learning for sparse coding, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, IEEE, 2012, pp. 3418–3425.
[56] F. Zhu, L. Shao, Weakly-supervised cross-domain dictionary learning for visual recognition, Int. J. Comput. Vis. 109 (1–2) (2014) 42–59.
[57] L. Shao, D. Wu, X. Li, Learning Deep and Wide: A Spectral Method for Learning Deep Networks.
[58] L. Shao, L. Liu, X. Li, Feature Learning for Image Classification via Multi-objective Genetic Programming.

Homa Foroughi received her bachelor degree from BuAli Sina University, Iran, in 2004 and the M.Sc. degree from Ferdowsi University, Iran, in 2009, both in computer engineering. She is currently working toward the Ph.D. degree in computing science at University of Alberta, Canada. Her research interests include computer vision, image processing and machine learning. She is a student member of IEEE.

Nilanjan Ray received his bachelor degree in mechanical engineering from Jadavpur University, Calcutta, India, in 1995, master's degree in computer science from the Indian Statistical Institute, Calcutta, in 1997, and Ph.D. in electrical engineering from the University of Virginia, Charlottesville, in May, 2003. After having 2 years of postdoctoral research and a year of industrial work experience, he joined the department of Computing Science, University of Alberta in July 2006 and now, he is an associate professor at Computing Science, University of Alberta. His research area is image and video analysis, segmentation, object detection, image classification and object tracking.

Hong Zhang received the B.Sc. degree from North-eastern University, Boston, USA, in 1982, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 1986, both in electrical engineering. He conducted post-doctoral research at the University of Pennsylvania from 1986 to 1987 before joining the Department of Computing Science, University of Alberta, Canada, where he is currently a Professor and Director of the Centre for Intelligent Mining Systems. He is an NSERC Industrial Research Chair. His current research interests include robotics, computer vision, and image processing.

Please cite this article as: H. Foroughi, et al., Robust people counting using sparse repregcisentation and random projection, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.02.009