Pattern Recognition 48 (2015) 3038–3052

Contents lists available at ScienceDirect

Pattern Recognition

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

Robust people counting using sparse representation 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
Available online 5 March 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:

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