Compressing Sparse Feature Vectors using Random Ortho-Projections

[1]Esa Rahtu, [1,2]Mikko Salo, and [1]Janne Heikkilä
[1]*Machine Vision Group, University of Oulu, Finland*
[2]*Department of Mathematics and Statistics, University of Helsinki, Finland*
erahtu@ee.oulu.fi

Abstract

In this paper we investigate the usage of random ortho-projections in the compression of sparse feature vectors. The study is carried out by evaluating the compressed features in classification tasks instead of concentrating on reconstruction accuracy. In the random ortho-projection method, the mapping for the compression can be obtained without any further knowledge of the original features. This makes the approach favorable if training data is costly or impossible to obtain. The independence from the data also enables one to embed the compression scheme directly into the computation of the original features. Our study is inspired by the results in compressive sensing, which state that up to a certain compression ratio and with high probability, such projections result in no loss of information. In comparison to learning based compression, namely principal component analysis (PCA), the random projections resulted in comparable performance already at high compression ratios depending on the sparsity of the original features.

1. Introduction

As computer vision problems are getting very diverse, there is an increasing need for powerful ways to describe the variety of objects and appearances which come up in applications [1, 7]. In many cases this has resulted in high dimensional descriptor spaces. It is not uncommon that the dimensionality of the feature space for example in face recognition goes beyond 70 000 [1] and in object recognition over 5 000 [7]. At the same time these descriptors are increasingly finding new real-time applications in hand held devices, where storage and computing resources are significantly limited [5].

The problem of high dimensional feature vectors is well known and several possible solutions have been proposed. One approach has been to develop new descriptors that capture the essential properties of some effective method, but are low dimensional by design. Some examples of such features include uniform pattern LBP [9], speeded up robust features (SURF) [2], and compressed histogram of gradients (CHoG) [5]. While these methods have been successful, this approach is generally difficult, because in each individual case it requires intensive design effort and often there is no natural low dimensional formulation.

Another solution is to apply general dimensionality reduction methods like principal component analysis (PCA), independent component analysis (ICA), kernel PCA, and locally linear embedding (LLE) [11]. A common property of these methods is that they require a training phase in order to learn the mapping function. This requires training data, whose amount is related to the number of dimensions in the compressed descriptor. In the case of very long vectors, where one is likely to need numerous dimensions in the compressed vector, it can be time consuming and expensive to acquire the needed representative training samples. In addition the lack of canonical mapping requires one to store the original features in addition to the compressed ones if these are intended to be used in multiple tasks. This is particularly inconvenient for large multipurpose datasets.

In this paper we investigate dimensionality reduction of sparse descriptors using ideas from compressive sensing theory [4]. In particular we apply projections onto a set of random orthogonal vectors as our compression method. The theory in [4] states that up to certain compression ratio and with high probability, such projections contain all information from the original features. The maximum compression, allowing perfect reconstruction, depends on the sparsity of the features [4], but our experiments indicate that the compressed descriptors perform very well in applications also far below this theoretical limit.

By sparsity we mean that feature vectors have large values only in a small subset of the elements, which may be different for each vector. More generally, the theory also applies to feature vectors which are sparse in some orthonormal basis. Sparse features are common

in computer vision methods like [1, 7, 3]. In [3] random projections are used in a special case of signature features, but in addition to that we are not aware of further applications of this technique to vision problems such as the ones presented here.

In this paper we will extend the study of random ortho-projection compression from the case illustrated in [3] into several common computer vision problems, namely texture classification, face recognition, and category recognition. The results are compared with specially designed low dimensional features and principal component analysis, which is by far the most popular learning based dimensionality reduction method. Our experiments show that random projections result in comparable performance already for high compression ratios, depending on the sparsity of the original features. The face recognition experiment also demonstrates the problems with training data in PCA.

2. Compressive Sensing

In this section we briefly describe those theoretical aspects of compressive sensing which are relevant for this paper. We will follow the survey [4] and refer to that article for further information and references.

In compressive sensing, one considers data which are assumed to be sparse in some representation. The goal is to find a way of compressing (or sampling) the data so that the original data can often be reconstructed from just a few samples. The point is that if the representations which achieve sparsity and compression are 'incoherent', one expects good reconstructions from fewer samples than required by classical results such as the Nyquist sampling theorem.

The data can be taken to be a vector f in $\mathbf{R}^n$. Consider two orthonormal bases $\{\psi_1, \ldots, \psi_n\}$ (the sparse basis) and $\{\varphi_1, \ldots, \varphi_n\}$ (the compression basis) in $\mathbf{R}^n$, and write Φ and Ψ for the orthogonal $n \times n$ matrices with rows φ_j^T and columns ψ_j, respectively. We say that f has an S-sparse representation in the basis Ψ if $f = \Psi x$ where at most S of the coordinates in $x = (x_1, \ldots, x_n)^T$ are nonzero (that is, x is S-sparse).

One would like to compress the vector f by computing inner products $\langle f, \varphi_k \rangle$ with vectors in the compression basis, and by selecting $y = (y_1, \ldots, y_m)^T$ to consist of some subset of m of these values. In terms of matrices, this can be written as $y = \Phi' f$ where Φ' is an $m \times n$ matrix obtained from Φ by selecting some m rows (in practice these can be selected randomly).

From the above relations, one can write $y = Ax$ where $A = \Phi'\Psi$. It follows from [4, Theorem 1] that if the bases Φ and Ψ are 'incoherent', one expects that from relatively few samples one gets a good approximation to a sparse vector x (thus a good reconstruction of the original signal f) by solving the ℓ^1-optimization problem

$$\min_{\tilde{x} \in \mathbf{R}^n} \|\tilde{x}\|_{\ell^1} \quad \text{subject to} \quad A\tilde{x} = y. \tag{1}$$

Such problems can be solved rather efficiently by convex optimization methods.

In the applications to computer vision, it is not always clear what a good sparsity basis Ψ could be. In fact, below we will use the identity matrix $\Psi = I$ as the sparsity basis (then sparsity means exactly that the vector f should have many components equal to zero). However, it follows from the theory (see [4, Section V]) that regardless of the choice of Ψ, one obtains with high probability a compression basis with the right incoherence properties by just taking Φ to be a random orthogonal matrix. Such random matrices provide a 'universal compression strategy' since one does not even need to know the sparsity basis Ψ to design a good data compression scheme.

We summarize the above discussion in the following (more precise) result, which follows from results in [4]:

Let Ψ be a fixed orthogonal $n \times n$ matrix, and let Φ' be an $m \times n$ matrix whose columns are obtained by orthonormalizing a set of m unit vectors in $\mathbf{R}^n$ which are chosen uniformly at random. Suppose that $f = \Psi x$ where x is S-sparse, and let $Ax = y$. If $m \geq C_0 S \log(n/S)$, then the solution x^* of the problem (1) satisfies $x^* = x$ with a high probability.

3. Random Ortho-projection Compression

In the applications below the feature vectors are sparse as given, and hence we have $\Psi = I$. The result in Section 2 then means that for sufficient sparsity and a reasonable amount of samples, the universal compression scheme based on random orthogonal vectors is lossless with high probability. Also, the original feature vectors can be reconstructed from the compressed information by solving (1).

Assume that f represents the original sparse feature vector in $\mathbf{R}^n$. Then the compressed features are achieved according to Section 2 as $y = \Phi' f$, where Φ' is an $m \times n$ matrix whose rows are orthogonal. In practice Φ' is constructed by applying the Gram-Schmidt orthonormalization procedure to a set of m randomly chosen unit vectors in $\mathbf{R}^n$. In addition if one has $\Psi \neq I$, then the compressed features are given by $y = \Phi'\Psi^T f$.

Finally, we note that since the compression is linear, it may be possible to embed the compression scheme into the computation of the original features. This can be done for instance in histogramming based methods.

(a) Outex (b) FRGC (c) Category database

Figure 1. Examples of the test images.

Since the elements of the feature vector correspond to different bins, one can directly compute the inner product of the feature vector with some row φ_j^T of Φ' by incrementing a running total according to the elements of φ_j^T instead of incrementing the elements of the feature vector by one. In this way one avoids having to store the original features. A similar approach was also used in [3] to achieve a fast and memory efficient implementation.

4. Experiments

We consider three common vision experiments including texture classification, face recognition, and category recognition. The results are compared with PCA compression, uniform pattern LBP, and original descriptors denoted as baseline. In texture and category recognition the PCA basis is learned from the same images that are used to train the classifier. In face recognition the learning is done using an extra training set according to [10]. The random ortho-projection compression is performed according to Section 3.

4.1 Texture Classification

We perform texture classification experiments using the publicly available texture recognition datasets Outex 00 and 02 [1]. The first set contains 480 128×128 images from 24 different texture types and the latter one 8832 32×32 images from the same texture types. Figure 1(a) shows some examples from the dataset. The original feature vectors were LBP [9] histograms, for which as an average 53.5% of the bins were nonzero and 80% of the sum was contained in 16.7% of the largest values.

The classification was performed using nearest neighbor classifier with L^2 distance and given 100 train-test-splits. The results are reported as average classification accuracy over all splits. Figures 2(a) and 2(b) contain the measured results with different descriptor lengths. The vertical lines represent the performances of LBP and uniform pattern LBP.

The results with Outex 02 show that already with 80 dimensions the compressed methods work almost as well as the baseline. For short lengths the learning based PCA performs better, but at about 100 dimensions random projections result in the same accuracy. Comparing to uniform pattern LBP, which has 59 dimensions, both PCA and compressive sensing seem to result in similar accuracy.

[1] http://www.outex.oulu.fi/

4.2 Face Recognition

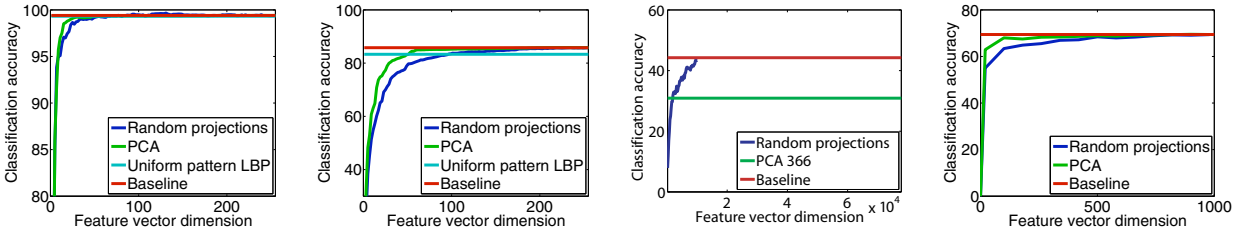
As a second experiment we ran Face Recognition Grand Challenge (FRGC) test 1.0.4 [10]. The test involves three image sets, one for additional training, one for training the classifier, and one for testing. These contained 366, 152, and 608 registered face images, respectively. Some examples are shown in Figure 1(b). The descriptors are constructed as presented in [1]: the preprocessed face image is divided into 304 non-overlapping regions of equal size and LBP histogram is computed from each of them. Finally, the obtained histograms are concatenated to form a 77 824 dimensional descriptor.

According to [10], the PCA basis can be learned using the additional training set. Since the basis is formed by the eigenvectors of an empirical covariance matrix, the maximum number of basis vectors is limited to 366. Furthermore the compression may also be sensitive to the actual training samples, since they occupy such a small portion of the feature space. This illustrates the difficulties encountered by learning based methods with very large dimensional feature vectors.

As an average only 13.0% of the vector elements were nonzero and 80% of the sum was contained in the 8.1% of the largest values. Hence these feature vectors are very sparse. The classification accuracies using nearest neighbour classifier with L^2 distance are shown in Figure 2(c). The results are computed up to 10 000 dimensions, where the random projections already achieve the baseline performance. We emphasize that no training was required in the compression, but it was enough to know that feature vectors are sparse.

4.3 Category Recognition

In the final experiment we applied compression methods to category recognition using the VOC 2007 dataset [6]. We chose to classify the "person" category,



(a) Outex00 (b) Outex02 (c) FRGC (d) Category recognition

Figure 2. Measured average classification accuracies. Notice the different scales.

since it contains the largest number of samples over the other classes, i.e. 1025, 983, and 2008 positive images in training, validation, and test sets respectively. Some examples from the dataset are shown in Figure 1(c).

The images were first converted to gray scale and resized to 320×240. We then extracted SIFT [8] on a regular grid with 10 pixel spacing and circular patches with radius 4, 8, and 12. The descriptors were vector quantized to 1000 visual words using K-means, and finally histogrammed over the words. As an average 73.3% of resulting 1000 bins were nonzero and 80% of the sum was contained in the 39.3% of the largest values, which makes these descriptors clearly the least sparse within the conducted experiments.

The classification was performed by training an SVM classifier with RBF-kernel, where the γ parameter was coarsely tuned using the validation set. Figure 2(d) contains the resulting average precision values for original and compressed descriptors. Results show that for very short lengths the learning based PCA performs better as expected, but interestingly already with 250 dimensions the difference to random ortho-projections is less than one percent.

5. Conclusions

In this paper we investigate the usage of random ortho-projections in the compression of sparse feature vectors. The method requires no further knowledge of the original features, and is therefore very attractive to applications where training data is costly or impossible to obtain. The independence from the data also enables one to embed the compression scheme into the computation of the original features in order to save time and memory. In the experiments we compared the approach to specifically designed low dimensional features and learning based compression, namely principal component analysis (PCA). The results indicate that already from relatively high compression ratios the random projection method achieved similar accuracy to PCA and even the original descriptors. This behavior was further emphasized with very sparse feature vectors.

References

[1] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkilä. Recognition of blurred faces using local phase quantization. *Proc. 19th Int. Conf. on Pattern Recognition*, 2008.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[3] M. Calonder, V. Lepetit, P. Fua, K. Konolige, J. Bowman, and P. Mihelich. Compact signatures for high-speed interest point description and matching. *IEEE International Conference on Computer Vision*, 2009.

[4] E. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, pages 21–30, March 2008.

[5] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod. Chog: Compressed histogram of gradients a low bit-rate feature descriptor. *IEEE Conference on Computer Vision and Pattern Recognition*, 0:2504–2511, 2009.

[6] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2007 (voc2007) results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[7] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:2169–2178, 2006.

[8] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[9] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002.

[10] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 947–954, 2005.

[11] L. van der Maaten, E. Postma, and H. van den Herik. Dimensionality reduction: A comparative review. *Tilburg University Technical Report*, 2009.