Resource-Allocating Codebook for Patch-based Face Recognition

Amirthalingam Ramanan and Mahesan Niranjan
School of Electronics and Computer Science
University of Southampton, SO17 1BJ, UK
{ar07r,mn}@ecs.soton.ac.uk

Abstract

In this paper we propose a novel approach to constructing a discriminant visual codebook in a simple and extremely fast way as a one-pass, that we call Resource-Allocating Codebook (RAC), inspired by the Resource Allocating Network (RAN) algorithms developed in the artificial neural networks literature. Unlike density preserving clustering, this approach retains data spread out more widely in the input space, thereby including rare low level features in the codebook. We show that the codebook constructed by the RAC technique outperforms the codebook constructed by K-means clustering in recognition performance and computation on two standard face databases, namely the AT&T and Yale faces, performed with SIFT features.

Keywords: Cluster analysis, Codebook, Face recognition, SIFT

1. Introduction

Face detection and recognition are active research areas that play important roles in many machine vision applications such as robotics, human machine interfaces, biometrics and surveillance. Though the performance of face recognition depends on a wide range of variation that includes pose, facial expression, illumination, occlusion, gender and time delay between acquisitions of image corpus, major advances have occurred in last decade. There are several known face recognition algorithms that make use of the information derived from the whole face, such as Eigenfaces (PCA) [1] and Fisherfaces (LDA) [2]. Another way to carry out face recognition is to use local feature-based techniques such as Local Binary Patterns [3] and SIFT [4].

The well known framework in the object recognition literature uses the SIFT descriptors to describe the patches and cluster them with K-means (KM) to generate a codebook that quantizes the features into vectors which are then fed in to a classifier as originally proposed in [5]. The classification performance of such an object recognition system depends on the efficiency of the visual codebook constructed by means of cluster analysis. Several other clustering techniques have been proposed in constructing codebooks for visual object recognition. Sivic *et al.* [6] used partitional KM clustering where as Mikolajczyk *et al.* [7] used a combination of KM and agglomerative clustering. Jurie and Triggs [8] used a mean-shift based clustering technique. Larlus and Jurie [9] used Gaussian Mixture Models (GMMs) as per-image clustering.

Clustering is a data transformation that preserves the distortion between cluster centres and the raw data. This need not to produce a discriminant codebook. There are several known difficulties with the use of KM clustering in this context including the choice of a suitable value for K, the computational cost of clustering when the dataset is large, and the convergence properties of the KM algorithm. In our experience the last of these is a major issue in defining the performance of an object recognition system. KM being an EM based algorithm converges to a local optimum close to the initial conditions.

In this paper we present a novel approach to designing a discriminant codebook that only processes each data item once, that we refer as a resource allocating codebook (RAC), inspired by the resource-allocating network (RAN) [10, 11]. A similar one-pass algorithm for inference from very large datasets has been developed in [12]. The RAN was developed as a means to overcome the problem of NP-completeness in learning fixed size networks, that can be used at any time in the learning process and the learning patterns do not have to be repeated. It either allocates a new unit, based on the novelty of a newly seen pattern, or adapts the network parameters by using the standard LMS gradient descent algorithm to fit that observation. The RAN can be interpreted from a function space approach to sequential learning. The aim of this paper is to evaluate the effectiveness of the RAC approach when applied to face recognition tasks. We demonstrate the discriminative power and computational efficiency of the RAC technique on two benchmark datasets the AT&T and Yale faces performed with SIFT features.

978-1-4244-4837-1/09/$25.00 ©2009 IEEE

The remainder of this paper is organised as follows. Section 2, presents the framework of our proposed algorithm together with its computational savings. Section 3, provides a brief description of our experimental setup with the testing results that support our claims. Also we show the robustness of the SIFT features with additive noise level. Finally, in Section 4 we conclude our paper.

2. Resource-Allocating codebook

Our proposed approach RAC, starts by arbitrarily assigning the first data item as an entry in the codebook. When a subsequent data item is processed, its minimum distance to all entries in the current codebook is computed, using an appropriate distance metric. If this distance is smaller than a predefined threshold, the current codebook is retained and no action is taken with respect to the processed data item. If the threshold is exceeded by the smallest distance to centroids, a new entry in the codebook is created by including the current data item as the additional entry. This process is continued until all data items are seen only once. Pseudo code for this approach is shown in Algorithm 1 below.

Algorithm 1 Resource-Allocating Codebook

```
Input: Visual descriptors ($\mathbf{D}$) and radius ($r$)
    of the hyperspheres.
Output: Centres of the hyperspheres ($\mathbf{C}$)
    Step 1: $C_1 \leftarrow D_1$
            $i \leftarrow j \leftarrow 2$
    Step 2: Repeat steps 3 to 4 until $i \leq$ size($\mathbf{D}$)
    Step 3: if min $\| \mathbf{D_i} - \mathbf{C} \|^2 \geq r^2$
            then create a new hypersphere of $r$ such that
              $C_j \leftarrow D_i$
              $j \leftarrow j + 1$
            endif
    Step 4: $i \leftarrow i + 1$
    Step 5: return $\mathbf{C}$
```

The novelty threshold used in RAC is regarded as a hyper parameter, and its choice has the same set of difficulties associated with the choice of *K* in KM. Our approach to set *r* is to take a small sample of the data, compute all pairwise distances between these samples and set the threshold, so that an approximate target codebook size is achieved.

Figure 1 shows a two dimensional projection of cluster centres found by KM and RAC techniques, projected on a plane defined by the first two principal components of the clustered data. While projecting from 128 to 2 dimensions masks much of the distribution, it can still be visualized that RAC gives codebook prototypes spanning in a wider range of space than KM.

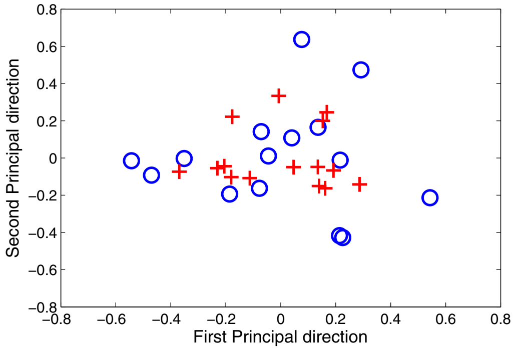

Figure 1. A comparison of visual codebooks designed by KM and RAC algorithms. '+' represents the cluster centroids of KM and 'o' represents the exemplars of RAC. For illustration purpose we plot the first two principal components of the cluster centroids obtained by KM (K=15) and RAC (*r*=1).

An illustrative example to compare the computational complexity of KM to that of RAC technique would be, $1559 \times \mathbb{R}^{128}$ SIFT descriptors were clustered into 160 clusters using KM in 19.79 seconds while RAC only needed 0.58 seconds to complete the one-pass execution on desktop computer with an Intel Core 2 running at 2.4GHz and 4GB of RAM.

The contribution of vocabulary entries together with the predefined threshold lead to a partitioning of the space into a set of overlapping hyperspheres when the distance metric used is the Euclidean norm. Local correlations between features could also be modelled in this framework by estimating covariance matrices associated with each vocabulary entry and using a Mahalanobis distance metric, similar to the sequential input space partitioning (SISP) algorithm in [13], though for simplicity we restrict ourselves to Euclidean distance in this paper.

3. Experimental work

3.1. Datasets

We tested our method on two benchmark face datasets. The first is AT&T (previously known as ORL) face database [14], containing 40 persons with 10 images per subjects. Images were taken at different times, varying the lighting, facial expressions and facial details. The second database is Yale face database [15], containing 15 persons with 11 images per subjects. Images were taken with different facial expression or configurations (see Figure 2).

We selected the testing image of each subject in a leave-one-out fashion and the remaining images for training. The number of training images per subject at each run was 9 and 10 for the AT&T and Yale faces, respectively. Since there are 40 subjects in the AT&T faces we totally obtained 40×9=360 training images and 40 testing images. In the

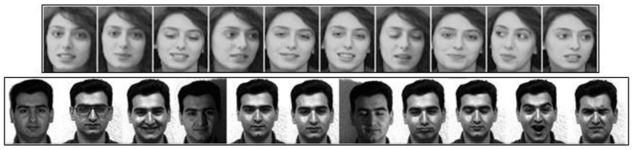

Figure 2. Example faces with different facial expressions: (top row) AT&T face with different poses, (bottom row) Yale faces with varying lighting conditions.

case of the Yale faces we totally obtained 15×10=150 training images and 15 testing images. This process is carried out until every image is used as a test image per subject. The reported results of our experiments are the average of these runs performed in a leave-one-out fashion.

3.2. Feature extraction

We used Scale-invariant feature transform (SIFT) [16], which is a representation of keypoints extracted from an image in a 128 dimensional space. We extract the SIFT descriptors automatically from all the images of the AT&T and Yale face databases without pre-processing the raw images and then used the KM and RAC methods independently to those descriptors extracted from the training images to construct the visual codebook. Codebooks are defined as the centres of the learnt clusters in the KM approach and the centres are selected in a resource allocating fashion from the actual descriptors in the RAC approach. Both training and testing images are then represented by the bag-of-keypoints approach by computing the frequency histograms with the codebook.

To check out the robustness of the SIFT features with additive noise level, we performed experiments with Gaussian noise by varying the standard deviation from 10 to 100 with increments of 10. Figure 3 shows an example subject in the AT&T faces with respect to the increase of noise level.

When we compare the performance versus noise level with the two face recognition tasks (see Figure 4), SIFT features perform better until σ = 30. Thereafter, the performance drops constantly with the increase of noise level. After σ = 50, the variations of performance in the Yale faces are very high in comparison with AT&T faces. This is because of the lighting variations and higher noise levels that make the recognition harder on the Yale faces.

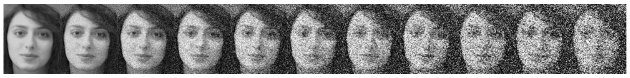

Figure 3. An example subject (left) in the AT&T face dataset. Left to right the standard deviation of the additive Gaussian noise varies from 10 to 100 with increments of 10.

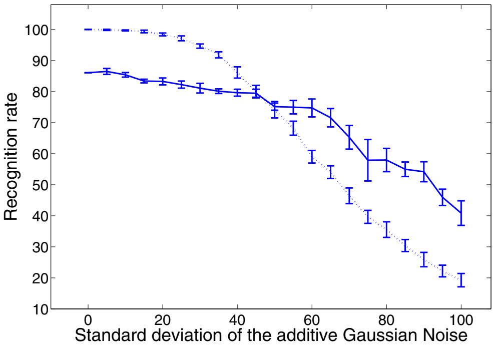

Figure 4. Recognition performance vs additive Gaussian noise (dotted line indicates AT&T faces and solid lines indicates Yale faces) over 10 independent runs.

3.3. Classification

The simplest classification approach would be a nearest neighbour voting strategy that computes all pairwise Euclidean distances between keypoint representation of a test subject to all labelled subjects in the dataset. This strategy is impractical in all but the smallest of problems. An alternative classification strategy, is to map the keypoints derived from an image in to a histogram of nearest cluster centre of a codebook, and then apply support vector machines (SVMs). SVMs are quite naturally designed to perform classification in high dimensional spaces. Classification in this paper was performed using a one-versus-all linear SVM. As a baseline, we also evaluate our approach with nearest neighbor (NN) classification.

3.4. Testing results

Results of classification experiments with KM and RAC are shown in Table 1. We used recognition rates as performance measures, which is the fraction of the correctly recognised faces to the total number of test faces. For the AT&T faces, the threshold of the RAC was *r*=0.7 and for the Yale faces, *r*=0.5 when the test image had the lighting variations, otherwise 1.0. In both face recognition tasks, KM uses K=1000.

SIFT features perform quite well and robust with different facial expression and pose, but fails to work under lighting variations [17]. SIFT features based on histograms of local orientation give some tolerance to illumination changes but not significant. Thus the performance drop in the Yale faces using SIFT features was due to the varying lighting conditions (see Figure 2, bottom row). To test the lighting effect with respect to SIFT features, we removed the images that has lighting variations (center-light, left-light and right-light) and tested the remaining images in a leave-one-out fashion. Table 1 (bottom row) shows the classification results.

Table 1. Recognition results on AT&T and Yale faces.

Database	NN	KM+SVM	RAC+SVM
AT&T faces	100%	98.25%	100%
Yale faces			
(with light effect)	86.06%	89.09%	95.15%
(without light effect)	100%	97.50%	100%

Both NN and RAC approaches carried out with SIFT features in a leave-one-out fashion outperforms the reported results in [18] on the AT&T faces (98.3%) and Yale faces (84.24%). The best quoted results in the literature on the Yale faces (100%) is by Branson and Agarwal [19] in which the images were cropped outside the face contour, aligned, Gabor filtered, z-scored and thereafter the dimensionality of the data is reduced by their proposed approach, structured principal component analysis (SPCA). They trained a perceptron by backpropagation on the training faces and tested on a novel face.

4. Conclusion

This paper shows how a discriminant visual codebook may be constructed by a resource-allocating algorithm. The approach takes one-pass through the data, making it computationally efficient. Because unlike clustering based algorithms, the retained codewords are not density preserving, they span a larger space retaining rare and discriminant features in the vocabulary set. Experimental results in the context of face recognition task performed with SIFT features, demonstrates the generality of our approach and the ease of implementation.

The greatest benefit of the proposed approach lies when applied on very-large datasets in visual object recognition tasks such as PASCAL Visual Object Classes Challenge[1] and Caltech-256[2], which we report elsewhere.

Acknowledgments

AR is supported in part by a grant from the University of Jaffna, Sri Lanka under the IRQUE Project funded by the World Bank and in part by a grant from the School of Electronics and Computer Science, University of Southampton, UK.

References

[1] M. Turk and A. Pentland. Eigenfaces for Recognition. In *Journal of Cognitive Neuroscience*, volume 3, pages 72–86, 1991.

[2] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos. Face Recognition using LDA-based Algorithms. In *IEEE Transactions on Neural Networks*, volume 14, pages 195–200, 2003.

[3] G. Zhang, X. Huang, S.Z. Li, Y. Wang, and X. Wu. Boosting Local Binary Pattern (LBP)-based Face Recognition. In *LNCS 3338*, pages 179–186, 2004.

[4] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli. On the use of SIFT Features for Face Authentication. In *IEEE Workshop on Biometrics, in association with CVPR*, 2006.

[5] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual Categorization with Bags of Keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[6] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, pages 1470–1478, 2003.

[7] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple Object Class Detection with a Generative Model. In *CVPR*, volume 1, pages 26–36, 2006.

[8] F. Jurie and B. Triggs. Creating Efficient Codebooks for Visual Recognition. In *ICCV*, pages 604–610, 2005.

[9] D. Larlus and F. Jurie. Latent Mixture Vocabularies for Object Categorization. In *BMVC*, pages 959–968, 2006.

[10] J. C. Platt. A Resource-Allocating Network for Function Interpolation. *Neural Computation*, 3:213–225, 1991.

[11] V. Kadirkamanathan and M. Niranjan. A Function Estimation Approach to Sequential Learning with Neural Networks. *Neural Computation*, 5:954–975, 1993.

[12] B. Farran and C. Saunders. Voted Spheres: An Online Fast Approach to Large Scale Learning. In *IEEE International Conference on Advanced Information Networking and Applications Workshop*, pages 744–749, 2009.

[13] R. S. Shadafan and M. Niranjan. A Dynamic Neural Network Architecture by Sequential Partitioning of the Input Space. *Neural Computation*, 6(6):1202–1222, 1994.

[14] F. S. Samaria and A. Harter. Parametrization of a Stochastic Model for Human Face Identification. In *IEEE Workshop on Applications of Computer Vision*, pages 138–142, 1994.

[15] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. In *PAMI*, volume 23, pages 643–660, 2001.

[16] D. Lowe. Distinctive Image Features from Scale-invariant Keypoints. In *IJCV*, volume 60, pages 91–110, 2004.

[17] J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, and B-L. Lu. Person-specific SIFT Features for Face Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 593–596, 2007.

[18] J. Yang, D. Zhang, A. Frangi, and J. Yang. Two-dimensional PCA: A New Approach to Appearance-based Face Representation and Recognition. In *PAMI*, volume 26, pages 131–137, 2004.

[19] K.M. Branson and S. Agarwal. Structured Principal Component Analysis. Technical report, UCSD, October 2003.

[1] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

[2] http://www.vision.caltech.edu/Image_Datasets/Caltech256/