Int J Comput Vis (2008) 76: 93–104
DOI 10.1007/s11263-007-0058-8

SHORT PAPER

Using Biologically Inspired Features for Face Processing

Ethan Meyers · Lior Wolf

Received: 4 March 2006 / Accepted: 2 April 2007 / Published online: 12 July 2007

Abstract In this paper, we show that a new set of visual features, derived from a feed-forward model of the primate visual object recognition pathway proposed by Riesenhuber and Poggio (R&P Model) (Nature Neurosci. 2(11):1019–1025, 1999) is capable of matching the performance of some of the best current representations for face identification and facial expression recognition. Previous work has shown that the Riesenhuber and Poggio Model features can achieve a high level of performance on object recognition tasks (Serre, T., et al. in IEEE Comput. Vis. Pattern Recognit. 2:994–1000, 2005). Here we modify the R&P model in order to create a new set of features useful for face identification and expression recognition. Results from tests on the FERET, ORL and AR datasets show that these features are capable of matching and sometimes outperforming other top visual features such as local binary patterns (Ahonen, T., et al. in 8th European Conference on Computer Vision, pp. 469–481, 2004) and histogram of gradient features (Dalal, N., Triggs, B. in International Conference on Computer Vision & Pattern Recognition, pp. 886–893, 2005). Having a model based on shared lower level features, and face and object recognition specific higher level features, is consistent with findings from electrophysiology and functional magnetic resonance imaging experiments. Thus, our model begins to address the complete recognition problem in a biologically plausible way.

Keywords Biologically motivated computer vision · Face identification · Face recognition · Learning distance measures · Kernel methods

E. Meyers (✉)
The Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
e-mail: emeyers@mit.edu

L. Wolf
School of Computer Science, Tel-Aviv University, P.O.B. 39040, Tel Aviv 69978, Israel
e-mail: wolf@cs.tau.ac.il

1 Introduction

Over the past ten years, appearance based approaches to face identification and object recognition have become increasingly popular. This rise has been supported by increase in machine learning methods that use statistical and probabilistic methods to discriminate between patterns found in natural images (Pontil and Verri 1998; Jones and Viola 2003). Yet it is becoming increasingly clear that the development new machine learning algorithms alone might not be the best approach to solving recognition problems, and that the feature set used has a large impact on the performance of these appearance based algorithms. Indeed, many of the best recognition algorithms in recent years have achieved their high levels of success due to new features sets (Lowe 2003; Ahonen et al. 2004). However, it still remains unclear what method should be used to derive new feature sets. Since the human visual system is the only example of a system that can perform identification tasks at a level of accuracy that is useful for most applications, it seems natural to try to understand and emulate how it represents visual data in order to derive features that will be useful in computer vision systems.

Within the field of computer vision recognition, object recognition and face identification are generally treated as separate problems with different feature representations, algorithms, and test sets used for each task. There are several

reasons why this division exists. The most obvious reason is one of practicality. If a system is designed to identify criminals, for example, having it report that it has recognized a zebra is a useless feature. The second, more interesting reason is that perhaps these tasks require different processing, feature representations or algorithms to be *effective*. Face identification requires the ability to detect subtle changes to a similarly shaped class of objects, while object recognition requires the ability to deal with large variations within an object class by finding specific indicative features, ignoring many other less relevant details. Despite these differences, many lower level problems, such as handling illumination changes, are common to both tasks. Moreover, many higher level classifiers that assign labels to given sets of features, are often similar among such systems. Reusing early representations and later stage classifiers offers not only a possible increase in efficiency, but also creates a highly versatile system that can be readily adapted to work with new visual tasks.

Literature from systems neuroscience shows that the primate visual system also uses a strategy of shared general early level processing that branches off into more specific higher level representations. Almost all visual information that reaches higher levels of the cortex first passes through center-surround processing in the retina and lateral geniculate nucleus (LGN), as well as through an early localized edge/spatial frequency representation in the primary visual cortex (V1) (Zigmond 1999). Later processing, specialized for identifying the location of an object, or recognizing what an object is, is done in different brain regions (Ungerleider and Mishkin 1982). Within regions involved in visual identification, there are areas that are more active for face images when compared to images of other objects (Kanwisher et al. 1997). Whether these areas are processing information in a fundamentally different way, or whether the same processing is being applied but different features are being used for faces and other objects, still remains an open question.

In this paper, we expand on the R&P model of object recognition (Riesenhuber and Poggio 1999) by building a new set of face-specific features. Similar to the strategy used in the visual system, shared early level representations are pooled in different combinations to build later stage face and object specific features. These face representations can then be run through the same types of classifiers, such as Nearest Neighbor Classifiers and Support Vector Machines to achieve a level of classification accuracy that matches some of the best algorithms currently available. By combining these results on face identification with the high performance previously shown for object recognition, the current implementation of the R&P model is starting to resemble a neurologically based computational system capable of approaching human performance on the 'complete' visual recognition problem.

2 Background

Below we discuss some advances in the field of face identification, and give a general background on the biological inspired image descriptors we use.

2.1 Face Identification Algorithms

Much progress has been made in the field of face identification and expression recognition. Commercial systems have been developed, and there are several standardized datasets to compare results against (Bolme et al. 2003; Phillips et al. 2002; Martinez and Benavente 1998). A few prominent approaches that have achieved a reasonable amount of success include applying Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) on the pixel intensities of the images (Etemad and Chellappa 1997; Turk and Pentlanbd 1991), as well as using a Bayesian Intra/Extrapersonal Classifier on the differences between pixel intensities of two images (Moghaddam et al. 1996). More recently, techniques using Local Binary Patterns (LBP) as features have outperformed these algorithms on the FERET and ORL datasets (Ahonen et al. 2004).

In this paper, we will primarily compare our results against LBPs and histograms of oriented gradient (HOG) representations. A comparison of our results to LBPs shows that our features can match the performance of one of the top representations used for face and expression identification. Comparing our results to histograms of oriented gradients illustrates that not all representations of oriented filters are equivalent. We also compare our results on the FERET database to PCA, LDA, and Bayesian Intra/Extrapersonal Classifier algorithms using the Colorado State University test system (Bolme et al. 2003).

Local Binary Patterns were originally introduced as a texture descriptor by Ojala (Ojala et al. 1996), and have subsequently been used as face and expression identification features (Ahonen et al. 2004; Shan et al. 2005). An LBP is created at a particular pixel location by thresholding the 3×3 neighborhood surrounding the pixel with the central pixel's intensity value, and treating the subsequent pattern as a binary number. A more general version of LBPs can be generated by specifying a particular radius around a central pixel, and a sampling value indicating the number of pixels to use from this radius. For example, specifying a radius of 2 and a sampling value of 16 (denoted (16, 2)), would consist of taking all the outer pixels in a 5×5 neighborhood surrounding the central pixels (see Ahonen et al. 2004 for more details). One further refinement consists of using only uniform binary patterns which are those binary patterns that have at most 2 transition from 0 to 1. For example, 1000111 is a uniform binary pattern while 1001010 is not. LBP representations for a given image are generated by dividing an

image into several windows and creating histograms of the LBPs within each window. In such representations, all non-uniform LBPs are treated as equivalent and given one histogram bin.

Histograms of Oriented Gradients (HOG) representations were introduced by Dalal and Triggs (2005) and have been shown to achieve some of the highest performance when used for pedestrian detection. Conceptually, these features are similar to the SIFT descriptor (Lowe 2003). First a histogram is defined wherein each bin represents a spatial location within the image as well as a particular orientation on the unit circle. In the published work the best results were obtained with one bin per 8 pixels in either spatial direction, and four or more orientations uniformly distributed from 0°–180°. Next, the magnitude and orientation of the brightness gradient is computed at each pixel. In order to calculate the feature vector, the magnitude of the gradients are added to the appropriate histogram bins by linearly interpolating between the nearest eight bin centers in both location and orientation. Finally, a normalization step is applied. In this paper we use a multi-scale version of these features by sub-sampling each image at half octave intervals over three octaves. Normalization is applied separately to each half octave. Preliminary tests on the ORL dataset (Samaria and Harter 1994) showed that this multiscale configuration achieves the best performance. It should be noted that HOG features have not been shown to be effective for face identification. We include them in our experiments so that we have a comparison against other front-end gradient based features.

2.2 R&P Model of Object Recognition

The R&P model of object recognition consists of several multi-scale hierarchical feed-forward layers of processing that correspond to different layers of processing found in the primate visual system. This model follows a similar structure as the classic Neocognitron model (Fukushima 1980), with alternating layers of simple (S) and complex (C) cell units creating increasing complexity as the layers progress from V1 to inferior temporal cortex (IT). One notable difference is that when pooling inputs at the C layers, a maximum operation over the S units is used rather than their sum.

The specific implementation of the model used for object recognition contains the following parameters. The first layer of the model, called the S1 layer, is created by convolving an array of Gabor filters at four orientations and 16 scales, over the input image. Pairs of S1 units at adjacent scales are then grouped together to create 8 'bands' of units for each of the orientations. The second layer, called the C1 layer, is then created by taking the maximum response within a local spatial neighborhood and across the scales within a band, to create a representation that contains 8 bands $\times$ 4 orientations. By taking the maximum filter response value within a small range of position and scale, tolerances to small shifts and changes in scale are built into the C1 representation. For the object recognition representations used in Serre et al. (2005), two higher level layers, called S2 and C2 layers, are built. In the S2 layer, conjunctive combinations of C1 units are learned from patches extracted from natural images. This combination of lower level features creates a higher level representation that is more selective and thus useful for discriminating between classes of objects. These S2 units are then convolved over an entire image and C2 units are assigned the maximum response value found from this convolution. Results from experiments using these C2 features achieve a high level of performance on a wide spectrum of object recognition tasks (Bieschi and Wolf 2005; Serre et al. 2005).

3 S2 Facial Features

The new set of facial identification features created in this paper, which we call S2 facial features (S2FF), contain three modifications to the R&P C2 features used for object recognition. The first modification consists of adding a center-surround stage of processing to each scale band prior to the extraction of S1 layer Gabor features. This processing is done by dividing the intensity value of each pixel by the mean of its intensity value and the intensity values of surrounding pixels within a window that is the size of the Gabor filter at a given scale. For example, if we are on the smallest scale where each Gabor filter occupies a 7 by 7 pixel region, then the intensity of each pixel is normalized by the mean value in a 7 by 7 region. From a biological perspective, this processing is analogous to the 'center-on surround-off' center-surround processing that occurs in the retina and in the lateral geniculate nucleus of the thalamus. From a computational perspective, center-surround preprocessing can eliminate intensity gradients due to shadows, and creates a representation similar to the self-quotient images that has been shown to be effective for face recognition (Wang et al. 2004). Results from our algorithm with and without this center-surround processing are reported in Sect. 5.

It should be noted that a more complete model of the retina and LGN would also include center-surround processing analogous to 'center-off surround-on' center-surround cells. We have decided to exclude this step of processing for the results reported in this paper since adding this processing doubles the number of features and hence memory and computational time required to run our algorithm. However, if 'center-off surround-on' processing is included by inverting the 'center-on surround-off process' (i.e., for each pixel, dividing the surround by the pixel's intensity), we notice a marginal improvement in performance (see Sect. 5.1).

Table 1 Parameters used to create the S2FF features. Filter size indicates the size of the filter used for the center-surround and the Gabor filters

	1	2	3	4	5	6	7	8
filter size s	7 & 9	11 & 13	15 & 17	19 & 21	23 & 25	27 & 29	31 & 33	35 & 37
rid size N^{Σ}	8	10	12	14	16	18	20	22
σ	2.8 & 3.6	4.5 & 5.4	6.3 & 7.3	8.2 & 9.2	10.2 & 11.3	12.3 & 13.4	14.6 & 15.8	17.0 & 18.2
λ	3.5 & 4.6	5.6 & 6.8	7.9 & 9.1	10.3 & 11.5	12.7 & 14.1	15.4 & 16.8	18.2 & 19.7	21.2 & 22.8
θ	$0; \pi/4; \pi/2; 3\pi/4$							

The second modification to the R&P model was to create face identification specific features using a linear combination of the C1 outputs. Weights for this linear combination were derived from a training set of images using a kernelized and regularized version of the relevant component analysis algorithm (Bar-Hillel et al. 2005) (KR-RCA) that is capable of dealing with high dimensional data. While using the KR-RCA algorithm to find the weights may not be biologically realistic, for the sake of showing that a linear combination of C1 features are capable of achieving a high level performance, the KR-RCA algorithm gives a convenient short-cut for finding appropriate weights. Until more evidence is gathered regarding the neurological mechanisms of perceptual learning, KR-RCA is a simple and effective surrogate.

One final difference that exists between S2FF features and C2 features concerns the final pooling stage. When calculating the final output value for a C2 feature, the feature is convolved over the whole image and the maximum output response is taken as the feature's value. This allows for large translational movements of object parts necessary to capture the high variability that is often seen within different exemplars of objects within a particular class. S2FF features do not perform a final maximizing stage, and thus these features are much more localized to a particular region in space. Such processing is better suited for identifying faces since faces have the same general shape and can be more accurately detected and aligned than most objects. Features from all scales and orientations are combined into one vector to form the S2FF feature set. The feature vector is subsequently normalized by having each data point sum to one and then taking the square root of each feature entry. Below is an algorithmic description of how to create S2FF features, and Table 1 lists the parameters that we used which are the same as were used by Serre et al. (2005). Figure 1 diagrams the construction of S2FF and C2 features.

1. Filter the image by applying center-surround divisive normalization processing. This is done by dividing each pixel's value by the mean value of the pixel and the pixels in it's neighborhood, where the neighborhood is of size $s \times s$. In our experiments s takes on 16 different values since we used 8 scale bands and we have two levels within each band.

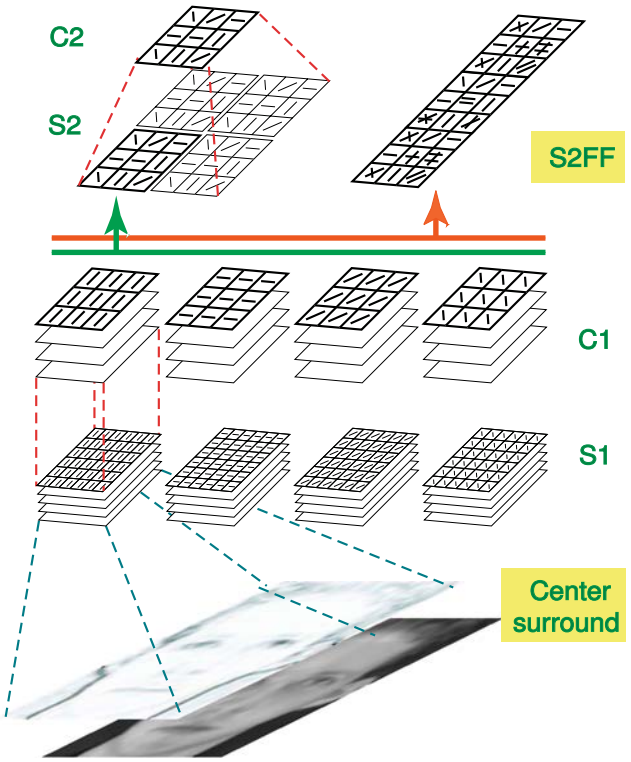

Fig. 1 A diagram showing how the S2FF and C2 features are created. S1 and C1 features are based on (Riesenhuber and Poggio 1999), S2 and C2 constitute an alternative pathway for object recognition as used in (Serre et al. 2005). Our contributions to the model are the addition of center-surround processing, and the S2FF features that are analogous to the face specific cells in the brain

2. Create the S1 representation by filtering each of the above center-surround images with Gabor patches that are the same size s that were used for the center-surround filtering, at four different orientations Θ. The equation for creating the Gabor filters is: $G(x, y) = \exp(-(X^2 + \gamma^2 Y^2)/2\sigma^2) \times \cos(2\pi X/\gamma)$, where $X = x\cos\theta + y\sin\theta$ and $Y = -x\sin\theta + y\cos\theta$.
3. Create the C1 representations by first taking the maximum value within a local neighborhood of size $N^{\Sigma} \times N^{\Sigma}$, where Σ is an index for a particular scale band. This creates a down-sampled representation. Then for each neighborhood, take the a maximum value within each scale band.

4. Apply KR-RCA to a concatenated list of all the C1 features across all orientations and scales to get the S2FF features.
5. Normalizing the features by having the representation of each image sum of one, and then taking the square root of each feature before applying a classifier usually helps improve performance.

For images in the Feret database, which are 150×130 pixels, the S2FF representation has dimensionality d of 15 772 features. For ORL images, which are 112×92 pixels, there are 8556 features, and for AR images, which are 144×192 pixels, there are 22 204 features. Since the number of images N in the ORL, AR and Feret, databases are 400, 533, and 3368 respectively, the number of features forms a highly overcomplete basis for these images, which is a similar strategy that seems to be used by the primate visual system (Olshausen and Field 1997). While the dimensionality of these representations could be reduced by either choosing coarser parameters when using the R&P model, or through features selection methods, we instead choose to create a kernelized version of RCA that can assign a similarity score to representations in this high dimensional space. The creation of this kernelized version of RCA is described in the next section.

4 "Kernel Regularized Relevant Component Analysis"

There is little biological evidence to guide how to build high level image descriptors. While there are many existing algorithms that can extract meaningful information from data, only specific algorithms which satisfy the constraints listed below are suitable for our purposes.

One significant constraint is that the method should be multiple class, and not binary in nature. This means that most common feature selection methods are not appropriate. Another constraint is that we would like to build generic high level face descriptors, and not ones that need to be retrained whenever a new face is introduced into the database. Indeed, in several of our experiments, the individuals who the training images were generated from were distinct from the individuals used to create the set of probe images. While many object recognition researchers have noted a boost in performance when employing a multiclass SVM compared to the nearest neighbor method, SVMs are not appropriate here. In the FERET dataset, the gallery contains just one face image per person, making the *direct* use of SVM, AdaBoost and other commonly used classifiers unattractive (however, see (Jones and Viola 2003) for a clever use of a such classifiers for this problem).

The recent study of data driven metrics (Hastie and Tibshirani 1996; Bar-Hillel et al. 2005; Goldberger et al. 2004), that are used together with a nearest neighbor classifier, provides a suitable framework. These methods embed the data in a Euclidean space such that points that belong to the same class are ideally close to one another while having large distances to points from other classes. We tried several such methods on a variety of small data sets, and while it is hard to draw definite conclusions, Relevant Component Analysis (RCA) (Bar-Hillel et al. 2005) was found to perform consistently among the best. RCA is also simple conceptually: given some data points that are known to belong to the same cluster, e.g., they are known to originate from the same person, the RCA algorithm finds a linear embedding transformation that minimizes the distances between the points. RCA is very similar to Linear Discriminant Analysis (LDA), except that LDA also tries to maximize the between class covariance.

Let the points $\{x_{j1}, x_{j2}, \ldots, x_{jn_j}\}, x_{ji} \in \mathcal{R}^d$ belong to such a group that has the same label (termed "chunklet" in Bar-Hillel et al. 2005). In each group there are exactly n_j points, and N is the total number of points. Assume that there are n such groups, which correspond, in our experiments, to n individuals in the training set. The metric embedding is given (implicitly) as a positive-definite matrix B such that the new distance between two points x and x' is given by $((x-x')^\top B(x-x'))^{\frac{1}{2}}$. Let m_j be the center of the j chunklet, i.e., $m_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ji}$. The RCA method minimizes the distances between the points in each chunklet by solving the following optimization problem:

$$\min_B \frac{1}{N}\sum_{j=1}^{n}\sum_{i=1}^{n_j}(x_{ji}-m_j)^\top B(x_{ji}-m_j) \quad \text{s.t. } |B| \geq 1,$$

where the constraint on the determinant prevents B from shrinking to zero.

This minimization problem is easy to solve algebraically, and the solution is given (up to scale) as $B = \hat{C}^{-1}$ where:

$$\hat{C} = \frac{1}{N}\sum_{j=1}^{n}\sum_{i=1}^{n_j}(x_{ji}-m_j)(x_{ji}-m_j)^\top.$$

The RCA algorithm has several disadvantages that we aim to rectify below:

1. It requires the computation and inversion of a $d \times d$ matrix $\hat{C}$. For problems of the size we consider below, and many other feature sets used currently in vision, this might be difficult and time consuming on a standard hardware.
2. Due to the inversion operator, it is unstable. This means that very small perturbations to the data can alter the outcoming result B dramatically.
3. As is, it works only for linear kernels.

The authors of (Bar-Hillel et al. 2005) suggest to combine the RCA transformation with a dimensionality reduction step, to help stabilize the algorithm, thus partly solving problem 2. However, in our experiments the dimensionality reduction did not provide satisfactory results. One reason may be that such a solution ignores directions that in the data that were not expressed during the training stage (see also the comparison with the pseudo-inverse solution below, which can be seen as one particular example of dimensionality reduction). Instead we found that a better alternative was to add a regularization term to the matrix $\hat{C}$ prior to inversion. The computational problem (1) is solved by working in the dual space, or in other words, by kernelizing the algorithm. This also enables us to work with non-linear kernels, which potentiality makes the algorithm much stronger. In our experiments, however, we did not observe a large advantage to non-linear kernels.

Regularization. Maybe the simplest way to add stability to the algorithm is to add the demand that all the points that are identical in all d coordinates except for one coordinate will have a small distance between them. This can be enforced by adding d groups of the form $\{x_k, x_k + \lambda_0 e_k\}, k = 1, \ldots, d$, where x_k is an arbitrary vector, e_k is a vector of all zeros except for the kth coordinate which is one, and λ_0 is a small constant that determines the amount of regularization.

With these added groups, the original matrix $\hat{C}$ would become the following regularized matrix:

$$\hat{C}_\lambda = \hat{C} + \sum_{k=1}^{d} \frac{\lambda_0^2}{2} e_k e_k^\top = \hat{C} + \lambda I_d,$$

where $\lambda = \frac{\lambda_0^2}{2}$ and I is the identity matrix in $\mathcal{R}^d$. By applying the simple smoothness demand we therefore obtain the Tikhonov regularization, with a constant λ.

In a preliminary set of experiments, we found that the performance is steady across a wide range of regularization parameters (λ). For example, for the ORL dataset (Samaria and Harter 1994), there is a large plateau of performance at the level of 96–99% depending on the features. This plateau happens for λ values between 10^{-10} to 10^{-3} times the largest eigenvalue of the covariance matrix $\hat{C}$. However, by taking the regularization parameter to zero, which is equivalent to taking the pseudoinverse of $\hat{C}$, performance drops by 3–10% depending on the feature representation. Having a regularization parameter which is too high (more than a tenth the largest eigenvalue of the covariance matrix $\hat{C}$) reduces performance by 1–5%. Throughout our experiments in Sect. 5, we used the regularization value of 10^{-4} times the largest eigenvalue of $\hat{C}$.

Kernalization. Let us order all the training points as the columns $x_1, x_2, \ldots, x_N$ of the matrix $X \in \mathcal{R}^{d \times N}$ and let $c(k)$ be the chunklet indicator of point k, i.e., if point k belongs to group j then $c(k) = j$. Let $R_j \in \{0, 1\}^N$ be the point indicator for class j, i.e., $R_j(i) = 1 \Leftrightarrow c(i) = j$. Consider the matrix $\hat{C}$ above, and notice that it can be written as $\hat{C} = ZZ^\top$, where

$$Z = \left[x_1 - m_{c(1)} | x_2 - m_{c(2)} | \cdots | x_N - m_{c(N)}\right] = XM,$$

$$M = I_N - \sum_{j=1}^{n} \frac{1}{n_j} R_j R_j^\top.$$

Let $K = X^\top X$ be the kernel matrix. As in many other kernel methods (Schölkopf and Smola 2002), the matrix X itself can be of infinite dimension or computationally infeasible, but it is sufficient to assume that we can compute the dot product of two of its columns by employing some kernel function.

The dual of the matrix $\hat{C} = (XM)(MX^\top)$ is the matrix $\hat{K} = (MX^\top)(XM) = MKM$ which is a matrix of size $N \times N$. As we will show below, in order to compute the RCA transform it is sufficient to manipulate the matrix $\hat{K}$, avoiding the need to manipulate a $d \times d$ matrix.

Let $D(s)$ denote the diagonal matrix with the elements of the vector s along its diagonal. Consider the following three Singular Value Decompositions: $\hat{C} = U_c D(s_c) U_c^\top$, $\hat{K} = U D(s) U^\top$ and $Z = XM = U_z D(s_z) V_z^\top$. Since $\hat{C} = ZZ^\top$ and from the properties of SVD $U_c = U_z$, similarly, $\hat{K} = Z^\top Z$ and therefore $U = V_z$. For the same reasons $s_z(i) = \sqrt{s_c(i)} = \sqrt{s(i)}$, $i = 1..\min(d, N)$. The decomposition $XM = U_z D(s_z) V_z^\top$ also implies that

$$U_c = U_z = XMV_z D(s_z)^{-1} = XMU D(s)^{-\frac{1}{2}}.$$

Assume that the decompositions above are *thin* decomposition, i.e., that the matrices U, U_c, U_z, V_z have the minimal number of columns and that all the elements of the vectors s, s_c, s_z are nonzero. The thin SVD decomposition can be recovered from the usual SVD decomposition by dropping columns out of the matrices, and trimming the s vectors.

Let $\bar{U}_c$ be an orthogonal basis to the subspace orthogonal to the matrix U_c. Note that $I_d = U_c U_c^\top + \bar{U}_c \bar{U}_c^\top$. Below $s + \lambda$ means that we add the scalar λ to every element of the vector s.

$$\hat{C}_\lambda = \hat{C} + \lambda I_d = U_c D(s_c + \lambda) U_c^\top + \lambda \bar{U}_c \bar{U}_c^\top,$$

and therefore

$$\begin{aligned} B = (\hat{C}_\lambda)^{-1} &= U_c D(s_c + \lambda)^{-1} U_c^\top + \frac{1}{\lambda} \bar{U}_c \bar{U}_c^\top \\ &= XMU D(s)^{-\frac{1}{2}} D(s + \lambda)^{-1} D(s)^{-\frac{1}{2}} U^\top MX^\top \\ &\quad + \frac{1}{\lambda}(I_d - XMU D(s)^{-\frac{1}{2}} D(s)^{-\frac{1}{2}} U^\top MX^\top) \\ &= \frac{1}{\lambda} I_d + XMU(D(s)^{-1} D(s + \lambda)^{-1} \\ &\quad - D(\lambda s)^{-1}) U^\top MX^\top. \end{aligned}$$

Notice that in order to compute expressions of the form $x^\top Bx'$ it is required to compute the scalar $x^\top x'$ and the vectors $x^\top X$ and $X^\top x'$. These are readily computed using the kernel function.

This new form of RCA (which we call KR-RCA), can handle much higher dimensional feature spaces than the original RCA (Bar-Hillel et al. 2005). In our experiments, however, we did not observe a large advantage to non-linear kernels, so the results for all experiments shown below are based on a form of KR-RCA that uses a linear kernel.

5 Experiments

To test the performance of S2FF representation for face identification, the FERET and ORL datasets were used. Tests using the FERET dataset were done using the CSU Face Identification Evaluation System, which implements the FERET test for semi-automatic face recognition algorithms with a few minor modifications (Bolme et al. 2003; Phillips et al. 2002). The CSU system preprocesses the images by registering eye coordinates, cropping an elliptical mask to exclude non-face regions and then equalizing the histogram of gray level intensities. For each algorithm/representation tested, a distance matrix is generated that contains a measure of the similarity between all pairs of images in the dataset. These distance matrices are then used to test different probe and gallery image sets to evaluate the performance of various algorithms. In this paper we report the results of the permutation tool, which uses a subset of FERET that contains 160 unique subjects, each with 4 images. On each iteration of the test, one probe image and one gallery image is chosen for each of the 160 subjects, and the result is marked correct if the shortest distance from a given probe is to the gallery image of the same subject. The permutation test runs for 10 000 iteration and returns statistics including the mean recognition rate, the 95% confidence interval, and the probability that one algorithm outperforms another.

The CSU system comes with implementations of PCA, LDA, Bayesian Intra/Extrapersonal face identification algorithms whose performance we compare against our S2FF features. Results reported for the Bayesian Intra/Extra Personal Classifier used the maximum a posteriori estimate (Bayesian MAP), which gave a higher accuracy than using the maximum likelihood estimate. Results for the PCA eigenfaces used the Mahalobis cosine angle, which showed higher performance than using the Euclidean distance metric. We also compare our results with results obtained from using a multi-scale histogram of oriented gradients (HOG) representation similar to that used in (Dalal and Triggs 2005) for pedestrian detection as well as to results obtained using LBP. Parameters for the LBP features consisted of using uniform binary patterns with a (16, 2) radius neighborhood, a 18×21 window size, thus copying the parameters used in (Ahonen et al. 2004). KR-RCA was also always applied to the HOG features since it always improved the results for these features.

Two different training sets were used to evaluate the performance of different feature sets and to measure how effective KR-RCA is in increasing each feature set's performance. The first training set consisted of a subset of the 'CSU standard' training (called feret_training_x2.srt in the CSU system). Images in this subset come from the fa and dup1 splits of the FERET dataset, and this training set is used to train the PCA, LDA and Bayesian MAP algorithms in the CSU system. The second training set consisted of half the images from the fc split of the FERET dataset (images 1110–1206) and the corresponding images from the fa split. This subset, called subfc, was used in the experiments reported by Ahonen et al. in their paper showing that LBPs are good features for face identification (Ahonen et al. 2004).

The ORL dataset (Olivetti Research Laboratory, Cambridge) consists of 10 different images of 40 subjects (Samaria and Harter 1994). The images vary along several dimensions including facial expression, scale up to 10%, and tilting and rotating of the head up to approximately 20 degrees. Tests on this dataset were done by randomly choosing 5 images of each individual as probe images and 5 images as gallery images, for 100 random permutations. Mean accuracies for S2FF, multi-scale HOG features, and for uniform LBP using a (16, 2) radius and with 30×37 window size are calculated, again exactly copying the parameters used in Ahonen et al. (2004). When applying KR-RCA, the gallery images were used as a training set.

Tests on face identification under different lighting conditions were also done on the AR data set (Martinez and Benavente 1998). In these experiments, images under three illumination conditions, left, right and frontal lighting were used as probes, and the gallery of images consisted of one neutral lighting image. When KR-RCA was applied, all four images of half the individuals were used for training, and the probe images consisted of the three illumination images taken from the other half of the individuals who were not in the training set. The gallery images for this experiment consisted of a single image of each individual in the probe set under the neutral lighting conditions.

5.1 Results

Table 2 shows the results from tests on the FERET dataset for the permutation test and for the fb, fc, dup1 and dup2 splits of the data, using the CSU standard training set. Results from the CSU permutation test show that the S2FF representation achieve a statistically significant higher mean rank one identification rate than the other five representations shown in the table (P(S2FF>other features)>0.995).

Table 2 Percent correct identification on the FERET database using 'CSU standard training set' for training

	FB	FC	DUP1	DUP2	Mean	Lower	Upper
S2FF	91.5	95.4	75.9	72.2	82.5	77.5	86.9
HOG	89.8	43.3	65.1	55.1	72.8	67.5	78.1
LBP w/ KR-RCA	92.2	51.5	67.7	56.4	73.0	68.1	78.1
LBP original	93.3	41.2	60.9	49.6	75.3	70.6	80.0
Bayesian_MAP	81.7	35.1	50.8	29.9	72	66.9	77.5
LDA	71.7	43.2	45.3	18.9	68.9	63.1	77.4
PCA_MahCosine	85.1	66.0	44.0	21.8	72.2	66.2	77.5

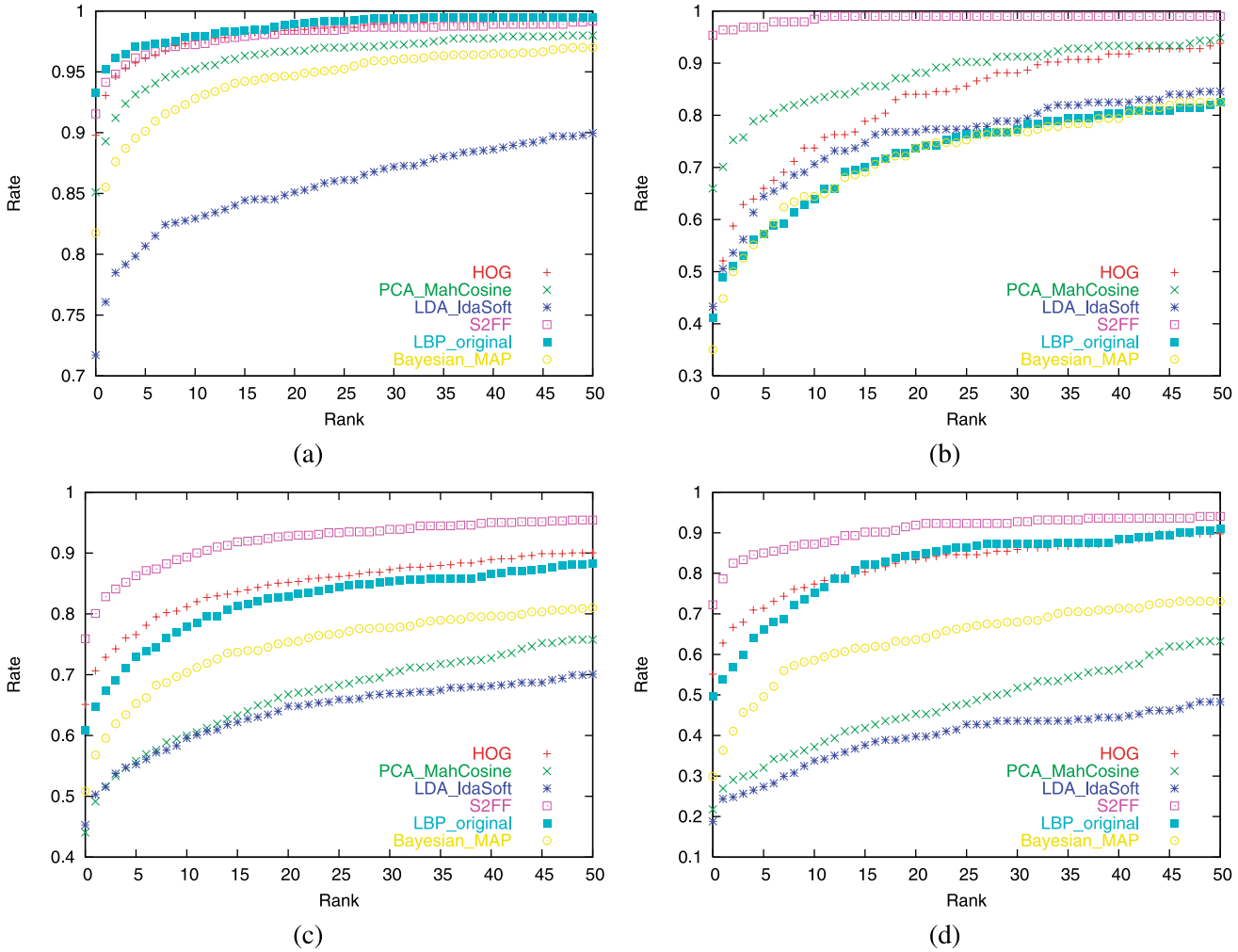

Fig. 2 Rank curves comparing S2FF, LBP, HOG, LDA, PCA Bayesian MAP face identification accuracies on different splits from the FERET database (**a**) FB, (**b**) FC, (**c**) DUP1, (**d**) DUP2. Training for the different representations was done using the CSU standard training set. As can be seen, S2FF features have a higher performance than other features on most splits. Image were generated by CSU Face Identification Evaluation System (Bolme et al. 2003)

Figure 2 shows the rank curves for the different representations.

Results on the FERET dataset using the subfc training set (as used in Ahonen et al. 2004) are shown in Table 3. Here S2FF features again receives a higher mean score on the permutation test than LBP and HOG representations although the difference between S2FF and LBP is not statistically significant at the alpha = 0.05 level - P(S2FF > LBP) = 0.89. One additional method used in Ahonen et al. (2004) to boost the performance of LBPs was to add a set of weights

Table 3 Percent correct identification on the FERET database using subfc split for training

	FB	FC	DUP1	DUP2	Mean	Lower	Upper
S2FF	90.5	96.4	56.9	65.0	81.9	76.9	86.2
HOG	90.0	74.2	54.0	46.6	75.8	70.6	80.6
LBP with KR-RCA	91.1	76.8	52.9	41.5	78.7	73.8	83.1
LBP and weights	96.7	68.0	64.3	59.4	79.5	75	83.8
LBP weights and KR-RCA	94.1	84.0	55.5	51.3	82.7	78.1	86.9
Combined S2FF and LBP	95.0	91.2	61.6	58.5	84	79.4	88.1

Table 4 Percent correct on face identification using the ORL and AR datasets

	ORL w/o KR-RCA (L2)		ORL with KR-RCA		AR w/o KR-RCA		AR with KR-RCA	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
S2FF	96.2	1.4	99.1	0.8	97.7	1.2	99.8	0.4
LBP	97.7	1.2	99.1	0.9	96.1	1.1	98.9	0.7
HOG	95.5	1.4	98.7	1.3	96.7	1.3	98.4	1.0

Table 5 Comparing different contribution to the S2FF performance on face identification (FERET mean results using the CSU permutation test)

Variant of SS2FF	FERET	ORL	AR
Original C1 features - No center-surround or KR-RCA	74.5	95.5	96.9
Center-surround, but no KR-RCA	74.9	96.2	97.7
No center-surround but with KR-RCA (subfc training)	76.1	98.7	99.2
No center-surround but with KR-RCA (CSU standard training)	74.5	N/A	N/A
Center-surround, KR-RCA, but no MAX (subfc training)	77.5	96.8	97.6
Center-surround, KR-RCA, but no MAX (CSU standard training)	52.7	N/A	N/A
S2FF CSU standard training	82.5	99.1	99.8
S2FF subfc training	81.9	N/A	N/A
S2FF_average CSU standard training	84.2	99.2	99.8
S2FF_average subfc training	82.1	N/A	N/A

to the histograms derived from different windows in the image before computing the chi-squared distance measure between images. The weights for each window were calculated using the subfc training set and classifying image using only one window at a time. The windows whose classification rate was below the 0.2 percentile received a weight of 0, windows whose rate was above the 0.8 and 0.9 percentiles received weights of 2 and 4 respectively. Results using this weighting are also shown in Table 3. Here we see that once the weights are added, weighted LBP with KR-RCA does achieve a higher mean score on the permutation test than the S2FF features although the difference is not statistically significant P(KR-RCA weighted LBP > S2FF) = 0.8072. When the weighted LBP features and S2FF are combined into the same vector and KR-RCA is applied, the mean performance is even higher. These results, however, are still not significantly different from the S2FF results P(LBP + S2FF + KR − RCA > S2FF) = 0.7885. The fact that combining LBP and S2FF achieves better performance than either alone, suggests that there is some unique information that each feature is capturing.

It should be noted that the LBP weights, which were obtained from (Ahonen et al. 2004) can lead to overfitting because the weighting scheme was designed heuristically, possibly to maximize performance on the testing set. Thus the results reported in Table 2 are probably the best indicator of LBPs true level of performance on the FERET dataset. Also, it is important to note that the original C1 and C2 features used by Serre et al. do not perform as well as the S2FF features created here (74.5% and 52.8% respectively vs. 82.5% for S2FF). A breakdown of which additions made to the C1 feature to create the S2FF features lead to an increase in performance can be seen in Table 5.

Table 6 Comparison of C1 and center surround processing with the original C1 features for object recognition. Each cell shows the area under the ROC curve of a classifier built using C2 features. The corresponding C2 features were computed based on the two types of the C1 features as in Serre et al. (2005)

Caltech dataset	C2	CS C2	CS C2 average instead of Max
Airplanes	95.4	98.9	94.2
Motorbikes	99.2	99.9	98.3
Faces	97.2	99.6	97.9
Cars	99.9	100.0	99.5
Leaves	99.0	99.4	97.3

Results from the ORL and AR face identification experiments are shown in Table 4. As can be seen in the table, LBPs and S2FF have the same level of performance on the ORL dataset while S2FF features have the best performance on the AR dataset.

Table 5 shows the performance of S2FF features when the center-surround processing, and the KR-RCA transform are left out. This table gives some insight into the relative contributions that components give to the overall S2FF performance. As can be seen, each modification by itself is only marginally beneficial, but the combination of the modifications creates a very significant improvement. This complex behavior, which we observed in other data sets and for other features, makes finding effective features a difficult task. We also looked at the affects of eliminating the maximum operator step at the C1 level, which is another way to compare our method to other techniques based on Gabor filter representations. To do this we used the smallest Gabor filter size in each band and we sub-sampled the S1 units over the same neighborhood sizes N^{Σ} used above but without taking the maximum value in this neighborhood first (using the larger size filters for each scale band gave similar results). As can be seen in Table 5, eliminating the application of the maximum operation also hurts performance.

Additionally, we tried replacing the max operation with an average operation by taking an average over the local neighborhood and between the two different sizes within each scale band (i.e., replacing the max operation with the average operation in step 3 of the algorithm listed in Sect. 3). Results from this modification were higher than the S2FF feature results on the Feret dataset, although not at a statistically significant level. Results between the average and max on the ORL and AR datasets were almost identical. However, changing the max operation to an average resulted in a decrement in object recognition performance due to the averaging eliminating the invariance to scale and position that the max operation gives (see Sect. 5.2).

Finally, we tested S2FF features with the addition of 'center-off surround-on' processing on the ORL and AR databases, and noticed a marginal increase in performance (99.97% and 99.94% respectively vs. 99.09% and 99.74% prior to adding these features). This increase, though small, made the S2FF features significantly better than the other algorithms on the ORL database. Comparing the S2FF with 'center-off surround-on' processing on the FERET database was not possible due to memory constraints.

5.2 Object Recognition Experiments

Since adding a center-surround stage of processing modifies the higher level C2 object representations as well as our new S2FF face representations, it is important to make sure that a high level of performance is still achieved on object recognition tasks. To verify this, we compared performance on the 5 CalTech datasets of Airplanes, Motorbikes, Faces, Cars and Leaves (Weber et al. 2000; Fergus et al. 2003; Fei-Fei et al. 2004). The center-surround operator was applied prior to creating the C1 representations, and then C2 features were created from these C1 outputs as was done in Serre et al. (2005). We used the original splits of the datasets, modified such that 30 negative examples were used for training, and removed from the testing set. A linear SVM was used for classification. In Table 6 we report the area under the ROC curve. As can be seen, adding the center-surround normalization does not hurt the performance of C2 features on object recognition tasks.

We also compared the performance of C2 units built on top of normal C1 units and C2 units built on top of center surround C1 units, on the multiple class problem of the 101 object dataset (Fei-Fei et al. 2004). A total 1000 C2 units of each type were computed in four different scales. The underlying prototypes needed to compute the C2 units were gathered from a set of "natural images". Using 15 training images per class and all the rest of the images as testing images, we got an average performance of 44.40% using the original C2 and 43.77% using the center surround units. In both cases the s.t.d was about 1%. Again, the differences in performance were not statistically significant. Finally we tested the C2 features that were created by taking the average response within a window and within each scale band instead of taking a max, and we noticed a decrease in performance.

6 Conclusions

In this work, we showed that at set of features based on what is known about the biology of the visual system is capable of achieving state of the art performance on face processing tasks. In order to construct these biologically inspired S2FF features, we modified a model of visual object recognition processing proposed by Risenhuber and Poggio, by adding center-surround processing to handle illumination changes, and introducing a new method of combining lower level features based on a kernelized and regularized version of the relevant component analysis transformation that is capable of handling high dimensional data. Tests on a several popular datasets showed that these S2FF features are indeed as good, and sometimes better, than other popular face image representations.

Since it was previously shown that R&P C2 features are capable of achieving some of the highest levels of performance on object recognition tasks (Serre et al. 2005; Mutch and Lowe 2006), and since our modifications to the R&P model do not disrupt the object recognition performance, our modified R&P model is beginning to address the 'complete recognition problem' in a biologically plausible way. Furthermore, using a set of shared set lower level features to deal with processing that is common to all visual tasks, as well as unique higher level descriptors to handle task specific functions not only emulates the processing of the human visual system, but it also is a much more efficient than having separate processing for all tasks. Finally, constructing a biologically plausible model of face and object recognition might not only be useful to the computer vision community, but it could also potentially provide insights psychologists and neuroscientists who are trying to understand the nature of how faces and objects are processed in the primate visual system.

Acknowledgements This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL).

This research was sponsored by grants from: DARPA Contract No. HR0011-04-1-0037, DARPA Contract No. FA8650-06-7632, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506, and National Institutes of Health (Conte) Contract No. 1 P20 MH66239-01A1. E.M. is supported by a National Defense Science and Engineering Graduate Fellowship.

Additional support was provided by: Central Research Institute of Electric Power Industry (CRIEPI), Daimler-Chrysler AG. Eastman Kodak Company, Honda Research Institute USA, Inc., Komatsu Ltd., Merrill-Lynch, NEC Fund, Oxygen, Siemens Corporate Research, Inc., Sony, Sumitomo Metal Industries, Toyota Motor Corporation, and the Eugene McDermott Foundation.

References

Ahonen, T., Hadid, A., & Pietikainen, M. (2004). Face recognition with local binary patterns. In *8th European conference on computer vision* (pp. 469–481).

Bar-Hillel, A., Hertz, T., Shental, N., & Weinshall, D. (2005). Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, *6*, 937–965.

Bieschi, S., & Wolf, L. (2005). A unified system for object detection, texture recognition and cotext analysis based on the standard model feature set. In *Proceedings of the British machine vision conference*.

Bolme, D. S., Beveridge, J. R., Teixeira, M., & Draper, B. A. (2003). The CSU face identification evaluation system: its purpose, features and structure. In *Conference on vision systems* (pp. 304–311).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Conference on computer vision and pattern recognition* (pp. 886–893).

Etemad, K., & Chellappa, R. (1997). Discriminant analysis for recognition of human face images. *Journal of Optical Society of America*, *14*, 1724–1733.

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *CVPR, workshop on generative-model based vision*.

Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Conference on computer vision and pattern recognition* (Vol. 2, pp. 264–271).

Fukushima, K. (1980). Neocognitron: a self organizing neural network model for a mechanism for pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Goldberger, J., Roweis, S., Hinton, G., & Salakhutdinov, R. (2004). Neighbourhood component analysis. *Neural Information Processing Systems*, *17*, 513–520.

Hastie, T., & Tibshirani, R. (1996). Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(6), 607–616.

Jones, M., & Viola, P. (2003). Face recognition using boosted local features. In *Proceedings of international conference on computer vision*.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–4311.

Lowe, D. G. (2003). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Martinez, A. M., & Benavente, R. (1998). *The AR face database*. CVC Technical Report #24.

Moghaddam, B., Nastar, C., & Pentland, A. (1996). A Bayesian similarity measure for direct image matching. In *Conference on computer vision and pattern recognition* (p. 638).

Mutch, J., & Lowe, D. (2006). Multiclass object recognition using sparse, localized features. In *Conference on computer vision and pattern recognition* (pp. 11–18).

Ojala, T., Pietikainen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, *29*, 51–59.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research*, *37*, 3311–3325.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2002). The FERET evaluation methodology for face recognition algorithms. *Pattern Analysis and Machine Intelligence*, *22*(10), 1090–1104.

Pontil, M., & Verri, A. (1998). Support vector machines for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(6), 637–646.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Samaria, F., & Harter, A. (1994). Parameterisation of a stochastic model for human face identification. In *2nd IEEE workshop on applications of computer vision*.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge: MIT.

Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. In *Conference on computer vision and pattern recognition* (Vol. 2, pp. 994–1000).

Shan, C., Gong, S., & McOwan, P. (2005). Conditional mutualinformation based boosting for facial expression recognition. In *Proceedings of the British machine vision conference*.

Turk, M., & Pentlanbd, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*, 71–86.

Ungerleider, L. G., & Mishkin, M. (1982) Two cortical visual systems. In: *Analysis of visual behavior*, (pp. 549–586). Cambridge: MIT.

Wang, H., Li, S., & Wang, Y. (2004). Face recognition under varying lighting conditions using self quotient image. In *IEEE international conference on automatic face and gesture recognition* (pp. 819–824).

Weber, M., Welling, M., & Perona, P. (2000). Unsupervised learning of models for recognition. In *European conference on computer vision* (pp. 19–32).

Zigmond, M. J. (1999). *Fundamental neuroscience* (1st ed.). New York: Academic Press.