Perceptual Distance Normalization for Appearance Detection

Chakra Chennubhotla & Allan Jepson
Department of Computer Science, University of Toronto
Email: {chakra,jepson}@cs.toronto.edu

Abstract

In this paper we develop a novel contrast-invariant appearance detection model. The goal is to classify object-specific images (e.g. face images) from generic background patches. The novel contribution of this paper is the design of a perceptual distortion measure for comparing the appearance of an object to its reconstruction from the principal subspace. We demonstrate our approach on two different datasets: separating eyes from non-eyes and classifying faces from non-faces. On the eye database, for a true detection rate of 95% we demonstrate a nine-fold improvement in the false positive rates over a previously reported detection model [5]. We also compare our detector model with a SVM classifier.

1 Introduction

In this paper we present a novel contrast-invariant appearance model for classifying object-specific ensembles (e.g. face images) from generic background patches. The contrast signal is the intensity variation around a mean brightness value. For the model proposed in this paper, we find it convenient to employ a generic, local contrast-normalization scheme based on the standard definition of Weber-contrast. For modelling appearance, we employ a linear, orthonormal basis from *sparse principal component analysis* (S-PCA) [4, 3]. The S-PCA basis have roughly similar reconstruction properties as PCA, but the PCA basis are global while S-PCA basis are sparse, spatially local, object-specific and multi-scale (that is, wavelet-like as shown in Fig. 1).

The novel contribution of this paper is the design of a perceptual distortion measure for comparing the appearance of an object to its reconstruction from a parameterized model, such as an S-PCA based principal subspace. It is well known that standard error norms, such as the mean-squared-error (MSE), are unsuitable for measuring perceptual distortion. Recent successes in formulating perceptual distortion norms (e.g. [13]) have come from analyzing the psychophysics of detecting spatially simple patterns, particularly contrast and orientation masking, and understanding the functional properties of neurons in the primary visual cortex. A typical perceptual distortion model consists of a linear transformation of images by a "hand-crafted" wavelet representation that is tuned to different spatial orientations and scales, followed by a divisive normalization mechanism. The normalization scheme involves pooling information in adjacent wavelet channels (that is in neighbouring spatial locations, orientations and scales, e.g., [2]). Normalization provides a context for local significance, in that a high sensor channel (wavelet) response is down-weighted if the adjacent channels are equally active but upgraded otherwise. The full set of normalized sensors tuned for different spatial positions, spatial frequencies, orientations and contrast discrimination bands provide a basis for assessing the perceptual similarity between two images (see [13] for more details).

Our work generalizes this normalization scheme to object-specific multi-scale representations derived from S-PCA. In particular, we show that after applying a linear model for the appearance and doing perceptual normalization, we can simply use the L_1 norm to separate the classes. For the task of separating images of eyes (extracted from the FERET database) from non-eyes, for a true detection rate of 95% we demonstrate a nine-fold improvement in the false positive rates over a previously reported detection model [5]. We also show results on the MIT face database [1].

2 Datasets

We investigate two different image datasets: eyes/non-eyes [5] and faces/non-faces [1]. The eye images are regions cropped from the FERET face database [6]. The face images were first scaled and rotated such that, in the warped image, the centers of left and right eyes have a horizontal separation of 40 pixels. From these warped images, we crop image regions around the eyes, each of size 20×25 pixels, to generate a database of 2392 eye patches.

For non-eyes, we construct a generic background patch ensemble by running an interest point detector [10] on several different natural images and collecting image patches with detector responses above a certain threshold. The interest point detector can be seen as a first step in the detection hierarchy in that it eliminates blank, texture-less regions from further consideration. To populate the 500 dimensional input space with a sufficiently large number of positive and negative examples, we symmetrize the ensemble. In particular, the eye images were flipped to generate mirror-symmetric pairs for a total of (2×2392) images. We take more liberties with the generic background patches, reflecting them about the x-/y-axis and around the origin, to make the original database of 3839 images 4 times as large. The datasets were randomly split in half to train and test the detection algorithm proposed here. We defer the details about the MIT face database to the results section.

3 Previous Work

In our previously reported detection method [5], we expand the test image in terms of an orthogonal basis derived from the training set. The orthogonal basis includes a constant DC image, the mean of the training set with the DC component removed and the leading M eigenbasis vectors. For a $M = 50$ dimensional PCA subspace, the true detection rate grows to 95% and the false positive rate reduces to $\approx 7\%$. We believe these recognition rates are less than satisfactory. A closer inspection of the eyes rejected as false negatives shows that these are images that look extreme, in that many have highlights caused by the specularities from the eye glasses or contain pixel outliers such as the hair falling over the eyebrows etc. It is possible to improve on the false negative rates by taking into account only relevant portions of an image in detecting eyes, as we have done in [5]. However, the high false positive rate remains major cause for concern and we present a vastly improved detection model next.

While we concentrate on appearance-based subspace methods, much work has been done in building feature-based object detectors [12, 11], in particular systems where the features are simple to compute and hence the objects are fast to detect [8, 7, 15, 9].

4 Detection Model

The problem of detecting known appearances is analogous to the problem of measuring image similarity, in that we need a perceptual error norm for meaningful results. Motivated by the work in [13], we propose a new detector model outlined in the figure shown to the right. There are five steps involved: (1) contrast-normalize ($\mathcal{WCN}$) the test image $\vec{x}$ to obtain $\vec{t}$; (2) project $\vec{t}$ into the wavelet-like space W derived from training S-PCA on generic background patches and obtain $\vec{d}$ as the coefficient vector; (3) build a low-dimensional approximation $\hat{\vec{d}}$ to the coefficient vector $\vec{d}$ using S-PCA basis B constructed for the object-specific ensemble in the "wavelet" space; (4) apply perceptual distance normalization $\mathcal{PDN}$ on the coefficient vector $\vec{d}$ and its reconstruction $\hat{\vec{d}}$ to obtain normalized vectors $\vec{z}$ and $\hat{\vec{z}}$; and finally (5) apply a simple detection strategy to $\vec{z}$ and $\hat{\vec{z}}$. We explain these details next.

$$\begin{tikzcd}
\vec{t} \arrow[d, "W^T"] & \vec{x} \arrow[l, "\mathcal{WCN}"'] & \\
\vec{d} \arrow[r, "B^T"] \arrow[d, "\mathcal{PDN}"] & \vec{b} \arrow[r, "B"] & \hat{\vec{d}} \arrow[d, "\mathcal{PDN}"'] \\
\vec{z} & & \hat{\vec{z}}
\end{tikzcd}$$

Step 1: Weber-Contrast Normalization ($\mathcal{WCN}$)

Weber-contrast is a measure of the relationship between the response of a pixel and that of its neighborhood. In particular, if x_i is the response of a pixel at location i and μ_i is an estimate of the mean response value in its neighborhood, then the Weber contrast signal c_i is defined as:

$$c_i = (x_i - \mu_i)/\mu_i. \tag{1}$$

The mean signal value μ_i can be obtained by convolving the image with a two-dimensional radially-symmetric Gaussian filter $G(i\,;\,\sigma)$. The neighborhood size is determined by the standard deviation σ of the Gaussian function. While this contrast computation removes shading variations, there are pixel outliers, such as the specularities from the eye glasses or the hair falling over the eye brows, that can bias the computation. To reduce the effect of outliers we normalize the contrast values using the following expression:

$$t_i = \frac{1-\exp(-\beta c_i)}{1+\exp(-\beta c_i)}, \tag{2}$$

where β is chosen such that for a predefined contrast value $c_i = c_{\text{def}}$, the normalized contrast t_i takes a value of 0.5. We set $\sigma = 3$ for estimating the contrast and $c_{\text{def}} = 0.3$ for normalization. In general, we observe larger values of σ improve the performance of the detector, but the detector is less susceptible to the actual setting of c_{def}.

Steps 2 & 3: S-PCA Representation: W, B

We use S-PCA over the standard PCA model for several reasons [4]. S-PCA is an orthonormal basis with directions rotated away from the PCA basis but with roughly similar reconstruction properties. The computation of S-PCA basis coefficients is efficient because of the presence of zero-valued weights in the basis vectors (see §3.6 in [3]). Finally, the S-PCA learning algorithm is simple and the optimization procedure is robust and scalable to high-dimensional spaces.

The S-PCA basis matrix trained on generic background patches is given by W, an $N \times N$ matrix. For an N-dimensional contrast-normalized image $\vec{t}$ the N-dimensional S-PCA coefficient vector is given by $\vec{d} = W^T\vec{t}$. Because the S-PCA basis look like wavelets, we abuse the notation slightly to call $\vec{d}$ a wavelet coefficient vector. Next, S-PCA is trained separately on the wavelet coefficients $\vec{d}$ generated for the images in the object-specific ensemble. For the following step, we build a low-dimensional representation for the wavelet coefficient vector $\vec{d}$ using the leading M object-specific S-PCA basis vectors. In particular, let B be the object-specific S-PCA basis matrix of size $N \times M$, then projecting the wavelet coefficient $\vec{d}$ gives $\vec{b} = B^T\vec{d}$ and the wavelet coefficient vector can be reconstructed as $\hat{\vec{d}} = B\vec{b} = BB^T\vec{d}$, which is again N-dimensional. Because the basis matrix B resides in the wavelet space populated by vectors $\vec{d}$ it is hard to interpret the basis vectors visually. Hence, in Fig. 1 we show the matrix $W * B$ obtained by pre-multiplying object-specific S-PCA basis B by the generic background patch S-PCA basis W. Notice, the basis $W * B$ is sparse, spatially local and multi-scale.

Step 4: Perceptual Distance Normalization ($\mathcal{PDN}$)

The coefficient vectors $\vec{d}$ and $\hat{\vec{d}}$ are now subjected to a perceptual distance normalization process. The idea is to normalize each wavelet coefficient by the pooled amplitude of wavelet coefficients tuned to similar spatial frequencies and similar spatial neighborhoods [13, 2]. Because S-PCA basis are learned from the data, as opposed to being hand-crafted, we need to find what the adjacent scales and orientations are

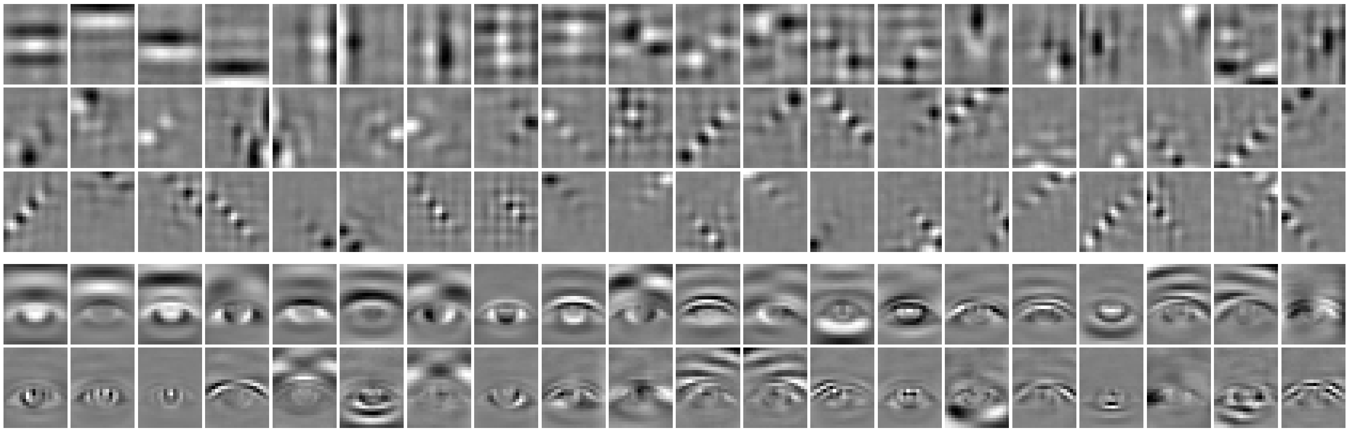

Figure 1: Leading multi-scale S-PCA basis for the datasets described in §2: (Rows 1-3) *Wavelets* for generic background image patches;(Rows 4-5) *Eyelets* for eye database. Notice the sparsity in the S-PCA basis, with zero value denoted by gray level 127.

for each S-PCA basis function in an automatic fashion. We outline a simple strategy next.

The spatial frequency tuning of the wavelets in W was observed to be related to the variance of the wavelet coefficients over the training set. Therefore by partitioning the variance spectrum we obtain subsets of wavelets tuned to roughly similar spatial frequencies. We automated this partitioning using a K-means clustering algorithm on the log of the variance spectrum. The result is shown in Fig. 2a, where the spectrum is partitioned into 5 groups, each drawn in a different color for the generic background patch ensemble.

To identify the local amplitude of image structure in each frequency band we form the amplitude image using a band of S-PCA basis indexed from l to h,

$$\vec{p} = \left| \sum_{l \le k \le h} \vec{w}_k d_k \right| . \qquad (3)$$

Here $\vec{w}_k$ denotes the k^{th} basis vector, d_k is the corresponding wavelet coefficient value. In Fig. 2b(TOP) we show an eye image and the bandpass images that result from the summation in the equation given above in each of the five different basis bands. In Fig. 2b(MIDDLE) we show the eye image along with its amplitude maps that result from the absolute operation over the bandpass images as given in the equation above. To provide a better intuition for the bandpass images, the images in Fig. 2b(BOTTOM) show the Fourier magnitudes of the bandpass images.

To estimate the portion of this amplitude image within the spatial support of the k^{th} wavelet $\vec{w}_k$, we compute:

$$s_k = |\vec{w}_k|^T \vec{p}. \qquad (4)$$

It can be shown that $s_k \ge |d_k|$ with the equality holding when $d_j = 0$ for $j \ne k$.

We can finally express the perceptual distance normalization ($\mathcal{PDN}$) of the k^{th} element of the coefficient vector as

$$z_k = d_k/(s_k + \upsilon_{lh}). \qquad (5)$$

The constant υ_{lh} is a saturation parameter for the basis band indexed from l to h. It is determined empirically by processing random images with a predetermined noise level (= 4 gray levels) and measuring the statistics of the resulting S-PCA coefficients. In particular, the random images are contrast normalized and for each wavelet band a corresponding amplitude map is generated. The amplitude maps are then projected back onto the wavelet space and the saturation constant υ_{lh} is set to the median value of the coefficients of the amplitude map in each wavelet band. The perceptual distance normalized coefficients of a wavelet coefficient vector $\vec{d}$ and its reconstruction $\vec{\hat{d}}$ are given by vectors $\vec{z}$ and $\vec{\hat{z}}$ respectively.

Step 5: Detection Strategy

For the purpose of detection we measure two numbers: (1) the wavelet norm given by the L_1 norm of $\vec{z}$; and (2) the error norm given by the L_1 norm of the error vector $\vec{z} - \vec{\hat{z}}$. We study the variation of these two numbers as a function of the increasing subspace dimensionality M, the number of columns in the basis matrix B. We expect the error norm to be high for generic image patches because the subspace was built for the object-specific ensemble. Also, we expect that the higher the wavelet norm, the higher will be the error norm for generic image patches. In fact, as we discuss next, what we observe is that the generic background patches and the object-specific ensemble appear as two distinguishable clouds with a small amount of overlap (Fig. 2c). We next present results using a straightforward detection strategy.

5 Results

Eyes/Non-Eyes: In Fig. 2c–d we show the results of applying the new detection method on the test set of eyes/non-eyes by varying $M = \{20, 50, 100, 200\}$. For clarity the plot has been scaled in such a way as to show all of the eye images (green points) at the expense of omitting a portion of the non-eye images (red points). As a detection strategy, we adopted a very simple approach of using a line aligned with the principal axis of the generic image patch cloud (red points). Points below the line are taken as positive detections. The ROC curve (Fig. 2d) is obtained by adjusting the y-intercept of the line. The ROC curve makes one thing very clear: the false posi-

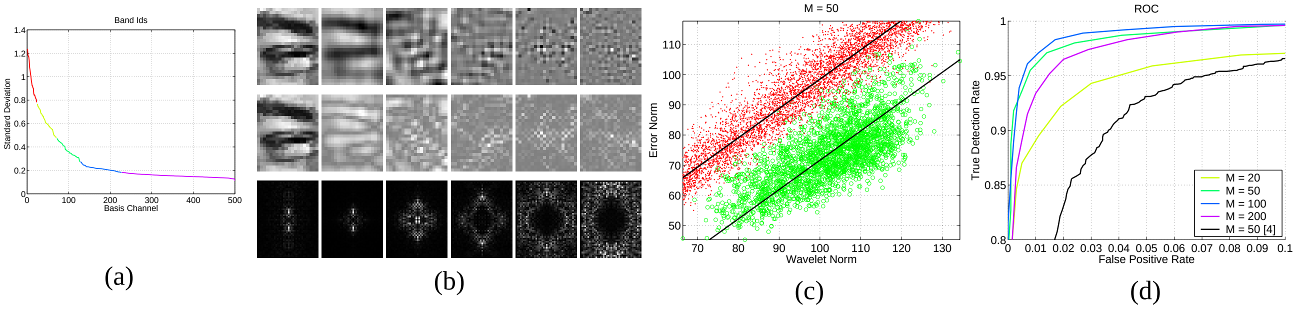

Figure 2: (a) Partitioning the variance spectrum into five regions. (b) An eye image and its bandpass components from the five different basis bands (TOP), the corresponding amplitude maps (MIDDLE) and the corresponding Fourier magnitude maps for the bandpass images (BOTTOM). (c) Characterizing eyes (green) and non-eyes (red) for the test set. (d) ROC curves for the eye test set using $\mathcal{PDN}$.

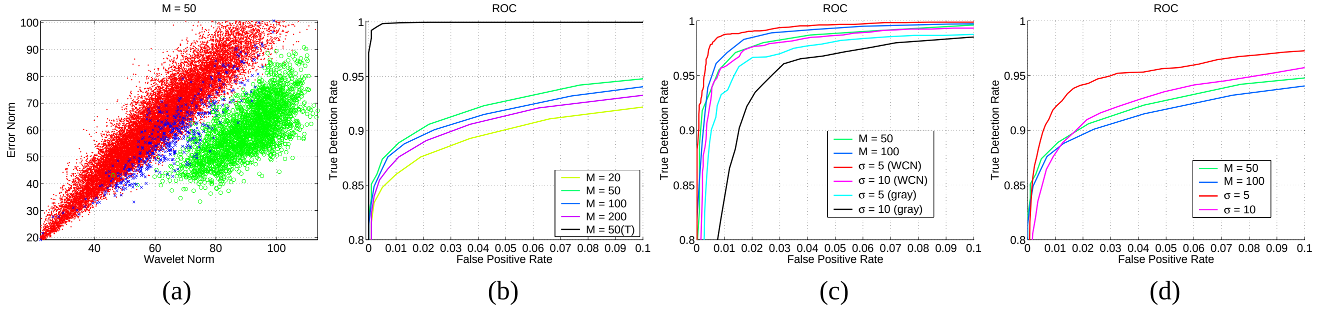

Figure 3: (a) Characterizing test-faces (blue), train-faces (green) (which together constitute the mixed-test set) and non-faces in the mixed-test set (red). (b) ROC curves for the "mixed" test set using $\mathcal{PDN}$. The black curve denotes recognition rate for just the train-faces in the "mixed" test set. (c & d) ROC curves comparing SVM with $\mathcal{PDN}$ for the eye test set in (c) and the "mixed" face test set in (d). The green and the blue curves correspond to $M = [50, 100]$ dimensional subspaces with the $\mathcal{PDN}$ model. The red and magenta curves show results from using SVM with $\sigma = [5, 10]$ on the contrast normalized ($\mathcal{WCN}$) dataset. The SVM performance on the unnormalized eye images (gray) is given by the black and cyan curves in (c). The $\mathcal{PDN}$ graphs for eyes and faces are identical to the ones shown in Fig. 2d and Fig. 3b.

tives can be kept very low, namely a value less than 0.8%, for a true acceptance rate of $\approx 95\%$ in a $M = 50$ dimensional subspace. This is a significant improvement over the previously reported detection model in [5] shown here in the same plot as a black line. In particular, for $M = 50$ the gain in false positive rate with $\mathcal{PDN}$ is nine-fold for a true detection rate of 95% and is 24-times better for a true detection rate of 90%.

Faces/Non-Faces: The MIT face database [1] consists of $2429/472$ training/testing face images and $4548/23573$ training/testing non-face images. Informally, most of the images in the training set are: cropped above the eyebrows, cropped just below the bottom lip, centered, roughly frontal views, and relatively well exposed. However, very few of the images in the test set appear to satisfy all these properties. We therefore created "mixed" training and testing sets by merging all these face images, adding the mirror symmetric versions, then randomly selecting half the dataset for mixed-training and the other half for mixed-testing. In Fig. 3a we plot the perceptually normalized space for the newly created testing set, where red dots indicate non-faces, green dots indicate faces from the original training set and blue dots indicate faces from the original testing set. Using $M = 50$ for a false positive rate of 0.1% we observe 16% false negatives, out of which 96% belong to the original testing set. In fact, the face images in the original testing set make up 16% of the mixed dataset and given the separation between the original training and original testing face images in the perceptually normalized space, this is not a surprise. Also, the black curve in Fig. 3b denotes the recognition rates obtained using $M = 50$ on just the original training set, while omitting the original test set. The recognition rates are near perfect.

We have also observed similar detection results using PCA models (instead of S-PCA) for object-specific and background ensembles with the same $\mathcal{PDN}$ formulation as provided in Eqs.$(3-5)$.

Comparison with SVM: We compare the performance of our detector with a publicly available implementation of a support vector machine (SVM) classifier [1]. SVM is parameterized by a kernel function and a C value which is the cost per unit violation of the classifier margin. We chose a Gaussian kernel and varied the σ parameter. The C value was set to 1, other values were tried but did not have a significant effect on the classifier.

On the eye dataset for different values of σ we observed a large variation in the total number of support vectors returned. In particular, for $\sigma = [3, 5, 10, 20]$ the number of support vectors returned on the training set are $[5267, 1564, 1140, 1616]$ respectively. Each support vector involves a dot product and hence, for a fair comparison the number of support vectors returned should be comparable to the number of inner products performed with the $\mathcal{PDN}$ model for a suitable choice of M. Thus, we selected $\sigma = [5, 10]$ and $M = [50, 100]$.

In Fig. 3c we compare the detection results from SVM to our $\mathcal{PDN}$ model on the eye dataset. The green and the blue curves denote the use of $M = [50, 100]$ dimensional subspaces with the $\mathcal{PDN}$ model. The $\mathcal{PDN}$ graphs are identical to the ones shown in Fig. 2d. The red and magenta curves show results from using SVM with $\sigma = [5, 10]$. The performance of SVM with $\sigma = 10$ is similar to using $\mathcal{PDN}$ with $M = 50$. Increasing the total number of support vectors, i.e. reducing σ from 10 to 5, improves the performance of SVM. In addition, we tested the performance of the SVM on the original gray-level images (i.e. without $\mathcal{WCN}$) (black and cyan curves in Fig. 3c). It is clear that contrast normalization causes a significant improvement in the performance of SVM.

We ran a similar experiment on the mixed training and testing sets that we created for the MIT face database. The number of support vectors obtained on the mixed training set for $\sigma = [5, 10]$ are $[1549, 1434]$ respectively. In Fig. 3d we compare the detection results from SVM to our $\mathcal{PDN}$ model on the mixed testing set. The green and the blue curves denote the use of $M = [50, 100]$ dimensional subspaces with the $\mathcal{PDN}$ model. The $\mathcal{PDN}$ graphs are identical to the ones shown in Fig. 3b. The red and magenta curves show results from using SVM with $\sigma = [5, 10]$. The performance of SVM with $\sigma = 10$ is similar to using $\mathcal{PDN}$ with $M = 50$. The best detection result we obtained was for $\sigma = 5$, which required 1549 support vectors.

A detailed comparison of the different methods is beyond the scope of this paper for several reasons: (1) The $\mathcal{PDN}$ normalization is *not* optimal in terms of computational efficiency. It was designed for simplicity. The normalization of each wavelet should depend only on a few “neighbouring” wavelets, and there is likely to be a more efficient way to do this than by generating the amplitude map; (2) It is not clear that the SVM implementation we have used is optimal (e.g. see [14]). If neither method is optimal, a detailed comparison may not be very revealing. Perhaps, most interesting is the use of several detectors (e.g., an eye, a nose, a mouth, and a face detector) within a single system. For such a system the wavelet transform used in our approach is common to all detectors, and hence the cost of the wavelet transform, in terms of the underlying hardware, can be amortized.

6 Conclusion

In this paper we developed a novel contrast-invariant appearance detection model to classify object-specific images from generic background patches. The novel contribution of this paper is the design of a perceptual distortion measure for comparing the appearance of an object to its reconstruction from the principal subspace. We showed an improved performance with our perceptual distance normalization ($\mathcal{PDN}$) based detection model. But this improvement comes with a price in that the images have to be represented in the full N-dimensional wavelet domain. However, we expect wavelet decomposition of signals to be a standard preprocessing tool. The simplicity of the detector in the wavelet domain is striking. In particular, after applying a linear model of eyes/faces and performing perceptual distance normalization we can simply use the L_1 norm to separate classes.

Ack: We thank G. Hinton and S. Roweis for discussions related to this work.

References

[1] M. Alvira and R. Rifkin. An Empirical Comparison of SNoW and SVMs for Face Detection. CBCL, MIT, A.I. Memo:2001-004, http://www.ai.mit.edu/projects/cbcl/software-datasets/FaceData2.html

[2] R W Buccigrossi and E P Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Tran. Image Processing*, 8(12):1688-1701, Dec 1999.

[3] C. S. Chennubhotla Spectral Methods for Multi-Scale Feature Extraction and Data Clustering. Ph.D Thesis. Dept of Computer Science, University of Toronto, 2004. http://www.cs.toronto.edu/~chakra/thesis.pdf

[4] C. Chennubhotla and A. Jepson. Sparse Principal Component Analysis. *ICCV*, pg. 641-647, 2001.

[5] C. Chennubhotla, A. Jepson and J. Midgley. Robust Contrast-Invariant EigenDetection. *ICPR*, 2002.

[6] Phillips P. J, H.Wechsler, J. Huang and P. Rauss The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing*, 16(5), 295-306, 1998.

[7] F. Fleuret and D. Geman. Coarse-to-Fine Face Detection. *IJCV*, 41(1/2):85–107, 2001.

[8] C. P. Papageorgiou and T. Poggio. A Trainable System for Object Detection. *IJCV*, 38(1):15–33, 2000.

[9] S. Agarwal and D. Roth. Learning a Sparse Representation for Object Detection. *ECCV*, 2002.

[10] Schmid, C. and R. Mohr. Local gray value invariants for image retreival. *PAMI*, 19(5):530-535, 1997.

[11] H. Schneiderman and T. Kanade. Object Detection Using the Statistics of Parts. *IJCV*, 2002.

[12] K. Sung and T. Poggio. Example-based learning for view-based detection. *PAMI*, 20:39–51, 1998.

[13] P. Teo and D. Heeger. Perceptual Image Distortion. *ICIP*, 2:982–986, 1994.

[14] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research, 1:211–244, 2001.

[15] P. Viola and M. Jones. Robust Real-Time Object Detection. *CVPR*, 2001.