Max–Planck–Institut für biologische Kybernetik
Max Planck Institute for Biological Cybernetics

Technical Report No. 119

Psychophysical Comparison of Synthesis Algorithms for Natural Images

Kristina Nielsen, Nikos K. Logothetis & Gregor Rainer

December 2003 .

This Technical Report has been approved by:

Director at MPIK

Postdoc at MPIK

Max–Planck–Institut für biologische Kybernetik
Max Planck Institute for Biological Cybernetics

Technical Report No. 119

Psychophysical Comparison of Synthesis Algorithms for Natural Images

Kristina Nielsen[1], Nikos K. Logothetis[1] & Gregor Rainer[1]

December 2003

[1] Department Logothetis, E–mail: kristina.nielsen@tuebingen.mpg.de, gregor.rainer@tuebingen.mpg.de

Psychophysical Comparison of Synthesis Algorithms for Natural Images

Abstract. In this study, we used three computational algorithms to compute basis sets for natural image patches, such that each patch could be synthesized as a linear combination of basis functions. The two biologically plausible algorithms non-negative matrix factorization (NMF) and sparsenet (SPN) were compared to standard principal component analysis (PCA). We assessed human psychophysical performance at identifying natural image patches synthesized using different basis set sizes in each of the algorithms. We also computed the reconstruction error, which represents a simple objective measure of synthesis performance. We found that the reconstruction error was a good predictor of human psychophysical performance. Performance was best for PCA, followed by NMF and SPN despite large differences in basis function characteristics. All algorithms were well able to generalize to represent novel natural image patches. When applied to white noise patches instead of natural images, PCA and SPN outperformed NMF. This shows that of the three algorithms the one that is least biologically plausible (PCA) actually supported best psychophysical performance, suggesting that in the present study it is low-level quality of reconstruction that is the main determinant of psychophysical performance.

1. Introduction

Recently, a new computational algorithm called non-negative matrix factorization (NMF) has been described which can learn a parts-based representation for complex visual stimuli such as faces (Lee & Seung 1999). It has been argued that this algorithm is biologically plausible, because it imposes non-negativity constraints on the contributions of each of the parts or basis functions (BF). A similar argument has been made previously for a different algorithm called sparsenet (SPN), which learns an efficient sparse code for the representation of natural images. SPN BF resemble the receptive field characteristics of primary visual cortical neurons (Olshausen & Field 1996). Both NMF and SPN can be used to find BF for a set of input images, such that each of these images can then be synthesized as a linear combination of the BF. The synthesized or reconstructed images can therefore only contain information that is expressible in terms of the BF. Since for biologically plausible algorithms BF capture image properties that are meaningful in terms of sensory coding of these images, images reconstructed with these algorithms should have higher information content for a human observer than images reconstructed with algorithms that are not biologically plausible. We thus hypothesized that a biologically plausible algorithm might be able to support better psychophysical performance than other arbitrary algorithms that are not biologically plausible. To test this, we compared the performance of NMF and SPN to principal component analysis (PCA) – a standard for finding BF that is not thought to approximate functions of the human visual system – using natural scene patches as input images.

In Experiment 1, human subjects were asked to identify patches synthesized with each of the algorithms in a delayed matching paradigm. Subjects saw an image synthesized using one of the three algorithms, and after a brief delay had to select the correct match from a set of eight images taken from the original set. By varying the number of BF (the basis set size) that each algorithm was allowed, we were able to assess the efficiency of each of the coding methods. In addition, we also compared the objective performance of each of the algorithms by computing the pixel-by-pixel mean square error between original and synthesized images. Experiment 2 was designed to test generalization performance of the algorithms. BF were computed for one set of images and subsequently used to represent images that had not been part of this set.

Experiment 1 demonstrated that both objective and human psychophysical performance improved with increasing basis set size, as expected. Regardless of basis set size, PCA performed better than NMF, and NMF performed better than SPN both in terms of reconstruction error and psychophysical performance. In fact, there was in general close agreement between reconstruction error and human psychophysical performance suggesting that at least for image patches at the same retinal position and size, the reconstruction error is a good predictor of human psychophysical performance. The ability to generalize was very similar for all three algorithms, as shown by the results of Experiment 2. Using BF to represent new images only had a small impact on the algorithms performance, both in terms of psychophysical performance and reconstruction error. Closer inspection of the reconstruction error revealed that the algorithms' generalization performance was worst for intermediate basis set sizes.

2. General Methods

2.1 Image sets

In the first two experiments, we used two sets of natural image patches, each consisting of 1000 pictures. These patches were drawn from different Corel Photo CDs, and showed natural motives like butterflies, flowers and various landscapes. No effort was made to exclude pictures with manmade objects, as long as they were only present in the background of the scenes. Originally, the images had a size of approximately 500 by 700 pixels. For further processing, a subregion of 256 by 256 pixels was selected in each image. This subregion was positioned randomly within the original picture such that its borders had a minimal distance of 100 pixels to the edges of the original image. The selected subregions were then converted to grayscale images. For natural images, the amplitude of the Fourier transform has a characteristic power law dependence (Field 1987). To ensure that all images employed here had identical Fourier amplitude spectra, we computed the Fourier amplitude of each image and then used the inverse Fourier transform with each image's Fourier phase spectrum together with the average Fourier amplitude of the entire set. After filtering, the images were rescaled to 16 by 16 pixels using nearest neighbor interpolation. This was necessary to allow the computation of BF in a reasonable amount of processing time. The two image sets were constructed in the same way and were similar in terms of mean image intensity and the distribution of grayscale values. A third image set consisting of 1000 random noise patch of size 16 by 16 pixels was generated with Matlab. Here, each pixel was set to a random value drawn from a Gaussian distribution with mean 0.0 and standard deviation 1.0. Subsequently, pixel values were rescaled to span the range from 0 to 1.

2.2 Calculation of basis functions

Algorithms to compute basis functions were implemented as Matlab code. For the two iterative processes NMF and SPN, the number of iterations was set such that the mean squared error between original and reconstructed images could converge to a stable value. About 4000 iterations were necessary in both cases to meet this criterion.

In all experiments, we varied the number of BF (the basis set size). Since PCA BF have a global order, restriction of the basis set size is possible by computing all BF at once and then selecting a subset of the appropriate size. BF are selected respecting their order, always starting with the first BF. For NMF and SPN, computation of BF was repeated for each basis set size.

2.3 Apparatus and Task

Stimuli were presented on a 21’’ monitor (Intergraph 21sd107) with a refresh rate of 85 Hz and a resolution of 1152 by 864 pixels. The monitor was gamma corrected for each channel individually. Background luminance was set to 17 cd/m^2. Subjects were comfortably seated in front of the screen with a viewing distance of 57 cm. The subjects’ head position was stabilized by a chinrest. A table in front of the subjects held a computer keyboard, which the subjects used for their responses.

Subjects performed a delayed matching-to-sample task (shown in Fig. 1), in which they had to match reconstructed images to their originals. Each trial started with the 500 ms long presentation of a reconstructed image (the target), followed by a blank screen of the background luminance for 50 ms. In the last frame, eight of the original images appeared and stayed on the screen until subjects made a response. All images had a size of 2.2° x 2.2°. The subjects had to indicate which of these images they had seen in the first frame. They responded by pressing one of eight keys on the keyboard in front of them. No limits were imposed on the reaction time. Once they had pressed a key, the next trial was initiated. There was no feedback about the correctness of their answer.

On each trial, the image used as target was randomly selected from the image set. Care was taken that each image appeared only once as a target within each session. Additionally, the method (e.g., the number of BF and the algorithm) used to reconstruct this image was randomized. The precise method depended on the experiment and thus will be described later. The subjects were given eight original images as response possibilities on each trial. To answer correctly, subjects had to select the image corresponding to the target. The difficulty of this task does not only depend on the similarity between the target and its original, but also on the differences between the eight original images to choose from. E.g., making these images very different may help the identification of the correct image even if target and original are quite dissimilar. To control for task difficulty, the seven original images representing incorrect answers were selected in dependence of the correct image on each trial. We calculated the mean squared difference, averaged over all pixels, between all images in the set and the correct match. According to this mean squared difference, the image set was split into three groups. Two images each were randomly chosen from the groups with the smallest and largest mean squared differences, respectively. Three images were selected from the medium group. The positions at which the original images appeared in the last frame were different in each trial.

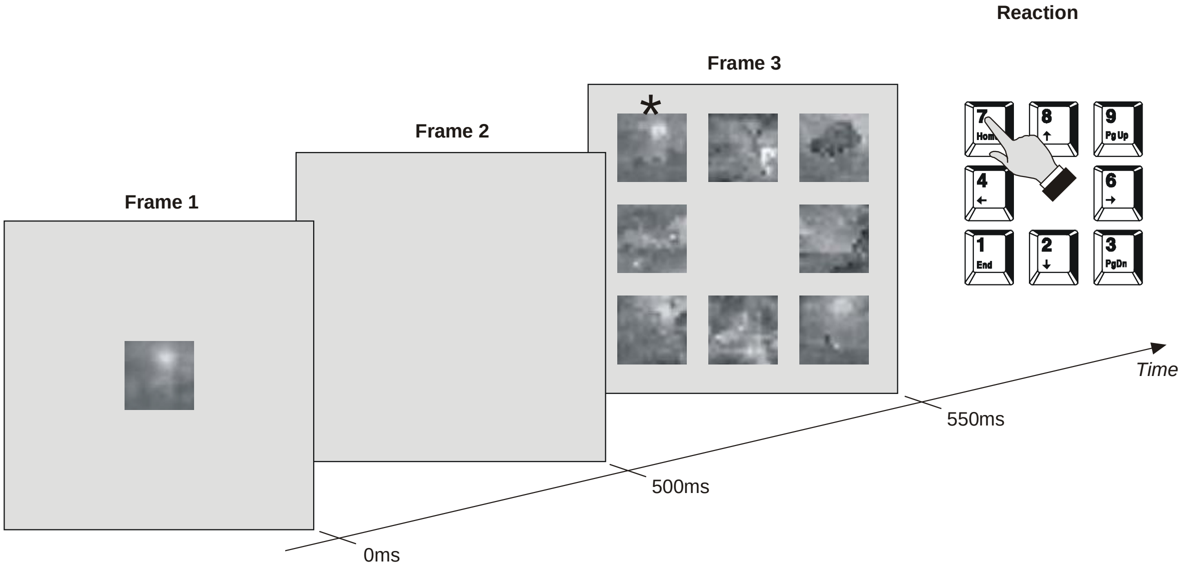

Figure 1 - Sequence of trial events: The target (a reconstructed image) was shown for 500 ms, followed by a blank for 50 ms. In the last frame, 8 images from the original image set appeared and stayed on the screen until the subject had given a response. The correct match, i.e. the original corresponding to the target, is indicated with a star (not shown in the actual experiment).

2.4 Reconstruction match

The difference between the reconstructed and original images was calculated as the mean squared error (MSE) between a reconstructed image and its corresponding original, averaged over all pixels and images. The MSE was normalized by the MSE between all original images, again averaged over all pixels and all possible combinations of original images. This normalized MSE approaches 0 for reconstructions that are very close to their originals. Differences between reconstruction and original as large as differences between original images lead to values close to 1. Thus, good reconstructions coincide with small normalized MSE values. To facilitate comparisons between analytical and psychophysical results, which were calculated as percent correct responses, we defined the reconstruction match (RM) as 1 minus the normalized MSE:

$$\mathrm{RM} = 1 - \frac{\frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{p=1}^{M}\left(r_{ip} - o_{ip}\right)^2}{\frac{1}{N^2}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{p=1}^{M}\left(o_{ip} - o_{jp}\right)^2}$$

where N is the number of images in the set, M the number of pixels per image, and o_{ip} and r_{ip} denote the value of the pth pixel in the original image i and its reconstruction, respectively.

3. Experiment 1

The core element of each of the three algorithms is a set of BF. These BF are extracted such as to capture specific properties of a set of input images. NMF and SPN apply biologically motivated criteria to compute the BF, while PCA employs a purely statistical criterion. Once BF are determined, the algorithms code input images as sets of BF activations. Because of this transformation, the algorithms can only represent information in the input images that is expressible in terms of the BF. The type and quality of this information ultimately determines the utility of an algorithm. The type of information can be determined by the properties of the BF. However, the quality of the information is more difficult to assess. We limited our analysis to one of its key features, namely the fidelity of the encoding.

Differences between input images and their BF representations can most easily be measured if both are represented in the same format. We therefore made use of the fact that all three algorithms are linear. Thus, the BF representation of an image can be transformed into a pixel-based representation – the reconstructed image – by computing a linear combination of the BF, weighing each BF by its activation (see Eq. 1 in the Appendix). Comparisons can then be carried out between a reconstructed image and its original.

We were especially interested in differences between biologically plausible and non-plausible algorithms. We hypothesized that since biologically plausible algorithms represent images by BF that resemble properties relevant for the sensory encoding of these images, differences between algorithms should be most notable if reconstructed images are compared to their originals by means of a psychophysical experiment. Biologically plausible algorithms should support better psychophysical performance than the biologically non-plausible ones. In this experiment, reconstructed images were compared to their originals by means of a delayed matching-to-sample task. Subjects were first shown a reconstructed image, and after a delay had to select the corresponding original out of a set of eight images. In addition to the psychophysical experiment, we also assessed differences between original and reconstructed images analytically, computing the reconstruction match as a pixel-based value for the mean squared difference between the images.

The number of extracted features is essentially arbitrary. For PCA and NMF, there exists an upper limit, which is the number of pixels of the original images. Between 1 and the maximum, all algorithms can extract any number of features. We varied the number of BF to assess its influence both on the extracted features and the reconstruction performance.

3.1 Methods

Basis functions

PCA, SPN and NMF computed BF for the same set of original images. Basis set size was set to 10, 25, 50, 100, 175 or 250. At the same time the coefficients for the images in this set were computed.

Procedure

The computed BF and coefficients were used to calculate reconstructions of the input images. Reconstructed images were computed separately for each condition, i.e. each combination of basis set size and algorithm. Subjects then had to match the reconstructed images to their corresponding originals in the described match-to-sample paradigm. There were 35 trials per condition, resulting in a total of 630 trials in the whole experiment, which usually lasted about 1 hour. Eleven naïve subjects participated in the experiment. For each subject, the number of correct responses per condition was computed. Finally, the data from single subjects was averaged to obtain the average number of correct responses per condition.

3.2 Results

Basis functions

Fig. 2A depicts the computed BF. Since each basis function is a function of pixels, they are represented as images. Comparison of the three panels in Fig. 2A illustrates the influences of basis set size. For a small set of BF, the functions were relatively similar over all algorithms, but they developed differently as basis set size increased. We considered as features parts of the BF with relatively homogenous values. Increasing the basis set size mainly resulted in a reduction of feature size, an effect that is especially pronounced for PCA and SPN. For the largest basis sets, BF of these algorithms hardly displayed any contiguous features. In contrast, NMF BF retained localized features even for the larger basis sets.

Feature size was determined quantitatively by analyzing the distribution of values in subregions of the BF. For this analysis, BF were again represented as images, and the standard deviation of BF values in square subregions of the BF was calculated as a measure for the homogeneity within this region. Both position and size of the subregion were varied. Position was moved such that each element of the BF served as upper left corner of the subregion once, under the constraint that the full subregion had to be placed within the BF. The subregion's size grew from 2 x 2 to 8 x 8 BF values. Standard deviations were averaged over all possible positions as a function of subregion size, and normalized by the standard deviation of all BF values. This ratio of standard deviations will be small while subregions are small and BF values within the subregion are relatively homogenous. It will increase with subregion size, as a consequence of more diverse values within larger subregions. The feature size of a BF was determined as the subregion size at which the averaged standard deviation in the subregions reached 80% of the standard deviation of the full BF (the critical size). To compute feature size for a set of BF, the critical sizes were independently determined for each of the BF and averaged over the whole set. Fig. 2B shows the analysis for a single BF (the PCA basis function Nr. 10). Its critical size was determined as 5 x 5 BF values. The left panel in Fig. 2B confirms that a square of this side includes just the white regions of this basis function. A comparison between the three algorithms is given by Fig. 2C. For small basis set sizes, critical size was large. With increasing set size, it decreased especially fast for PCA and somewhat slower for SPN. For a set of 250 BF, a critical size of less than 2 x 2 BF elements confirmed that these BF contain no features. In contrast, NMF feature size remained large even for large basis sets. It fell from a critical size of over 8 x 8 elements to about 6 x 6 elements for 250 BF.

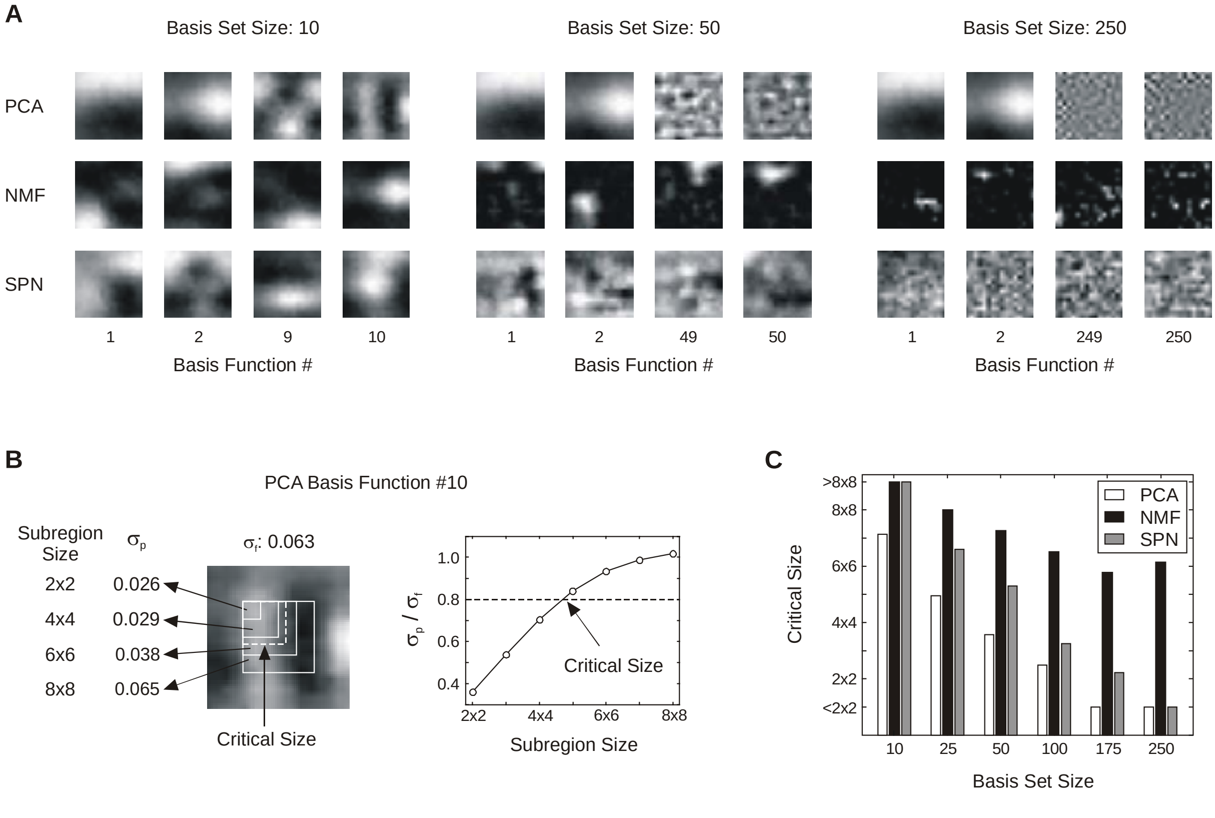

Figure 2 - Basis function characteristics: (A) Overview of the BF computed for PCA, NMF and SPN at three different basis set sizes. **(B)** Analysis of feature size for an example basis function. The left panel illustrates how different patch sizes influence the standard deviation of pixel values within these patches (σ_p), with large patch sizes leading to higher pixel variance. The center panel plots σ_p averaged over all possible patch locations and normalized by the standard deviation of all pixels in the entire basis function, σ_f. From this plot, the critical patch size was determined as 5 x 5 elements**. (C)** Dependency of critical patch size on algorithm and basis set size, illustrating that for small set sizes BF were similar across algorithms with differences developing with larger basis set sizes.

Reconstruction performance

The psychophysical experiment tested how well the different basis sets represented the input images as a function of basis set size. The average number of correct responses in each condition is plotted in Fig. 3A. All algorithms showed a similar trend, in that recognition performance increased with basis set size. Paired t-tests between all three possible pairs of algorithms, separately performed for each basis set size, revealed no significant differences either between PCA and NMF or between NMF and SPN at any basis set size. There were however significant differences between PCA and SPN for intermediate basis set sizes of 50 (t[10]=2.375, p=0.039) and 100 (t[10]=2.988, p=0.014).

The average number of correct responses increased asymptotically with basis set size. Changes in performance were only present until a certain set size had been reached. Thereafter, enlarging the basis set size left the performance unchanged. The minimal basis set size after which performance remained stable was determined by comparing the performances at adjacent basis set sizes with paired t-test, separately executed for each algorithm. In the case of PCA, the tests revealed significant changes in performance up to a set size of 50. NMF performance changed until 100 BF were used in the reconstructions. For SPN, dependence on basis set size was weaker. While there were significant differences between 10 vs. 25 and 50 vs. 100 BF, no further differences were seen at higher basis set sizes. The results of all t-tests are summarized in Table 1.

On each trial, subjects responded by selecting one of eight images. One of these images was the correct image, i.e. the original corresponding to the reconstructed image shown in this trial. The other seven images were selected such that two of them had a small distance to the correct image, three a medium and two a large distance. Distances were measured in terms of the MSE. A closer inspection of subjects' errors revealed that they preferentially selected images that fell into the first two categories (Fig. 3B). Averaged over all algorithms and basis set sizes, 80% of the errors involved images that either had a small or a medium distance to the correct image. Only in 20% of the trials subjects erroneously selected images that were very different from the correct image.

Table 1: For each of the algorithms, subjects' performances were compared between different basis set sizes using a t-test. The table lists the resultant p-values.

compared basis set sizes	PCA	NMF	SPN
10 vs. 25	.047	.016	.002
25 vs. 50	.009	.038	.067
50 vs. 100	.060	.004	.026
100 vs. 175	.156	.026	.060
175 vs. 250	.377	.369	.508

Calculation of the RM under the same conditions allowed a comparison of psychophysical and computational results. As Fig. 3C shows, there was a close correspondence between psychophysical and computational results. Similar to the psychophysical results, RM only slightly differed between algorithms. PCA led to the best RM, followed by NMF and SPN. The dependence of RM on the basis set size displayed the same asymptotical behavior as described for the performance of human observers.

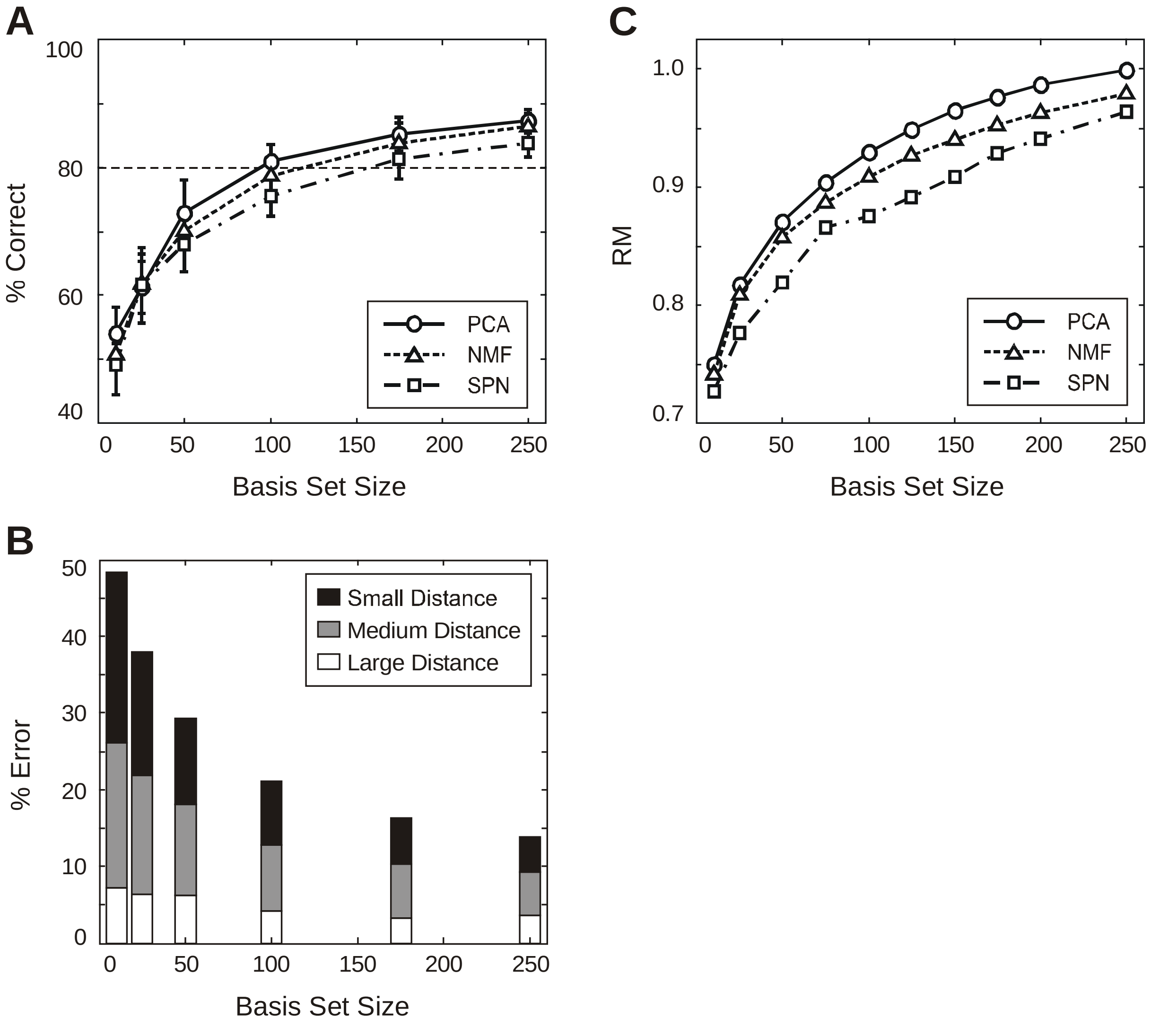

Figure 3 - Results of Experiment 1: (A,B) Psychophysical results: **(A)** Percentage of correct responses per algorithm and basis set size, averaged over all subjects. Error bars show the SEM. **(B)** Distribution of error types in dependence of basis set size. **(C)** Computational results: RM as a function of algorithm and basis set size.

4. Experiment 2

BF are derived by adaptation to a particular set of images. After the BF computation, any image can in principle be coded using the then fixed set of BF. The second experiment was designed to test how capable the algorithms were in representing "novel" images, i.e. how well they could generalize. We calculated BF for one set of images and used them to represent images from a second set. Again, reconstruction performance was assessed both psychophysically and analytically. Since basis set size may have an influence on the generalization ability, it was varied similar to Experiment 1.

4.1 Methods

Basis functions

Two image sets were used in this experiment. For both of them we separately computed BF. These BF were then applied either to reconstruct images from the set they had been computed from or from the other set, which was "new" to the BF. The first image set, i.e. the one that the BF were derived from, will be termed training set, the latter one transfer set. Since BF were computed for the two image sets, each image set could serve as training or as transfer set. Coefficients for the training set were calculated at the same time as the BF. To compute coefficients for the transfer set, the update rules given in the General Methods were used, while keeping the BF fixed. As in Experiment 1, the basis set size was varied. It was set to 10, 50, 100, 175 or 250 BF.

Procedure

For each subject, only one of the basis sets was used. Images both from the corresponding training and transfer set were reconstructed. Trials in which images from the training set were shown were grouped into a "direct" condition, whereas trials with images from the transfer set were selected for the "generalization" condition. For each combination of algorithms, reconstructed image set and basis set size, there were 22 trials. The implemented basis set and thus the image sets representing training and transfer set changed between subjects. A total of 20 naïve subjects participated in the experiment, ten for each basis set. The experiment lasted about 1.5 hours. For each of the subjects, the number of correct responses per combination of reconstructed image set, basis set size, and algorithm was registered. To calculate the average number of correct responses, we combined the results of both groups of subjects to exclude any influence of the basis set on the results.

4.2 Results

In the "direct" condition, images were reconstructed that belonged to the image set the BF were computed from. This and the range of basis set sizes tested was identical to the first experiment. Since two different basis sets were used, the second experiment extends the results of the first experiment. A comparison of the average number of correct responses between the two experiments showed the two experiments in good agreement. When BF were used to reconstruct images in the "generalization" condition, the performance of human observers decreased slightly. The difference in the average number of correct responses between "direct" and "generalization" condition, calculated independently for each subject and then averaged over all subjects, is plotted in Fig. 4A.

To test the influence of basis set size and reconstructed image set on the performance, we entered the number of correct responses into a repeated measures ANOVA with the within-subjects factors condition ("direct" vs. "generalization" condition) and basis set size. Basis set identity was used as a between-subject factor. ANOVAs were performed separately for each algorithm. For SPN, differences between "generalization" and "direct" condition were present at each basis set size (F(1,18) = 17.965, p<0.000). In contrast, PCA performance did not depend on condition. NMF had an intermediate performance. Here, a significant interaction between condition and basis set size was found (F(4,72)=3.078, p=0.021), which was due to a difference between "generalization" and "direct" condition at a set size of 175 BF (ANOVA at 175 BF, F(1,18)=7.067, p=0.013). There was no influence of basis set for any of the algorithms.

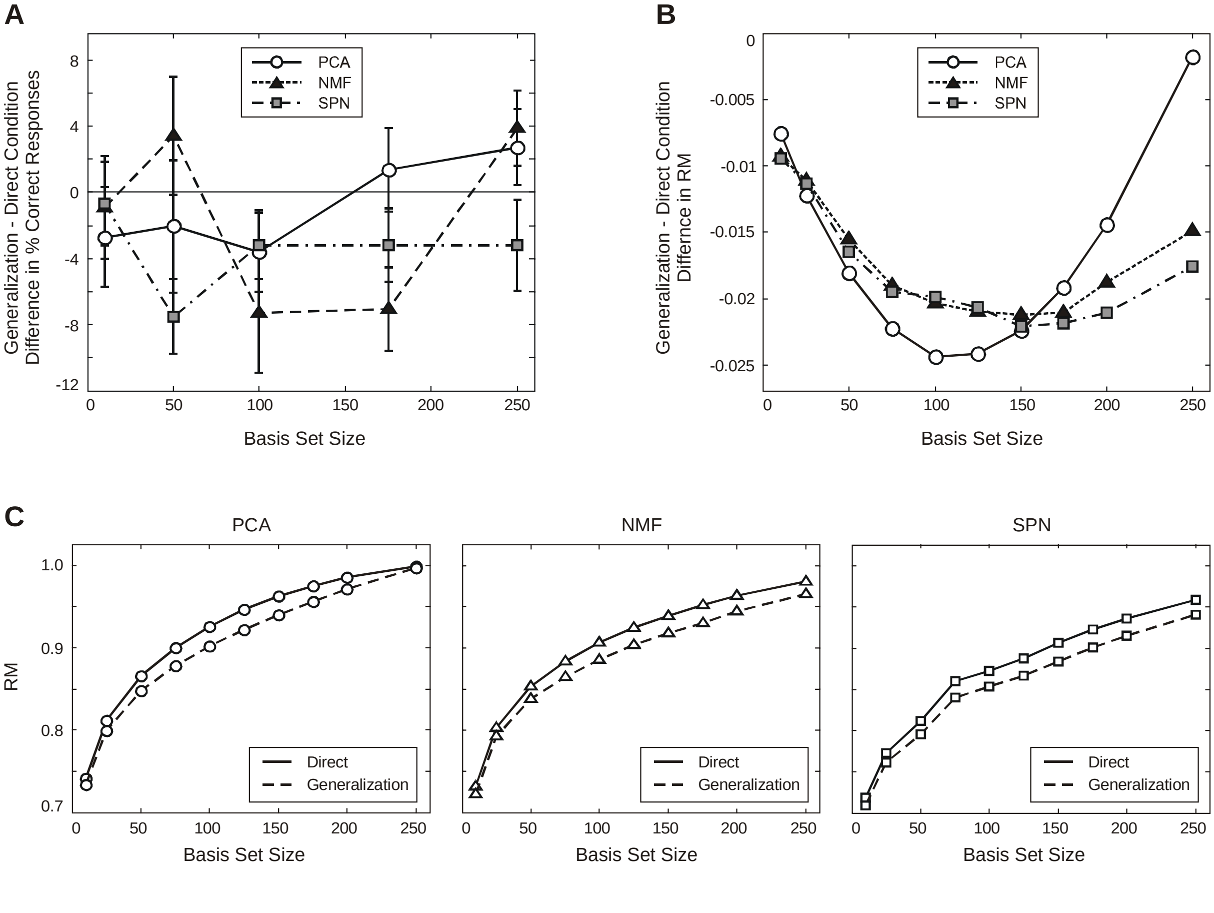

Figure 4 - Results of Experiment 2: (A) Psychophysical results: Differences in percent correct responses between "generalization" and "direct" condition, averaged over all subjects. Error bars depict the SEM. Negative values indicate a performance that is worse in the "generalization" condition than in the "direct" condition. **(B,C)** Computational results: **(B)** Difference in RM between "generalization" and "direct" condition. Negative values indicate a lower RM in the "generalization" than in the "direct" condition. **(C)** RM in the "generalization" and "direct" condition as a function of basis set size, separately plotted for each algorithm.

The RM was calculated using the two basis sets to reconstruct their respective training and transfer set. The results were averaged separately for the training and the transfer sets. Conditions are named as in the psychophysics experiment, with the "direct" and "generalization" condition referring to results derived from the reconstruction of the training and transfer set's images, respectively. A plot of RM in the two conditions (Fig. 4C) shows that generalization led to a decrease in reconstruction performance for all three algorithms. These changes in RM were small compared to the actual values of the RM.

To analyze the changes more thoroughly, we calculated the difference in RM (the generalization error) between the two conditions (Fig. 4B). All algorithms had a small generalization error for both small and large basis set sizes, with a maximal error at intermediate set sizes. Thus, small basis set sizes – although they allowed only a limited reconstruction performance – were relatively general. The increment in performance with increasing set size seemed to be generated by adapting the BF closely to the image set they were computed from. Only after a certain set size had been reached, a good reconstruction performance could be achieved with more general BF. The basis set size with the maximal error was smaller for PCA (100 BF) than for NMF (150 BF) and SPN (150 BF).

5. Experiment 3

The three algorithms were developed with different goals. NMF's update rules were set up such as to mimic certain biological principles, while SPN was adapted to encode natural images very efficiently. PCA in contrast is a more general purpose algorithm based only on sources of variance in the data. So far we have tested the algorithms' capability to encode natural scenes. Both experiments demonstrated almost equally good performances for all three algorithms. In the third experiment, we tested how well the algorithms represent "unnatural" input, namely images in which the distributions of pixel intensities followed a normal distribution. We hypothesized that if the principles underlying SPN and NMF make them specifically adapted to natural input, their performance should drop under these conditions. The reconstruction performance of PCA should however be relatively unaffected. As before, reconstruction performance was quantified psychophysically and analytically as a function of basis set size.

5.1 Methods

Basis Functions

An image set of 1000 white noise patches was generated as described in the General Methods section. BF were computed with NMF, PCA and SPN, using the same basis set sizes as in Experiment 1. Together with the BF, the coefficients for the patches were computed.

Procedure

As before, the computed BF, weighted by the respective coefficients, were combined to yield reconstructions of the input patches. These were then compared to the original patches in the DMS task. Eight subjects participated in the experiment, performing 35 trials per combination of basis set size and algorithm. The experiment usually lasted about one hour. Subjects' performance was analyzed by computing the number of correct responses for each combination of basis set size and algorithm. Group data was computed as the average over subjects.

5.2 Results

Using the algorithms to reconstruct white noise patches instead of natural scenes lead to an overall reduction in performance (see Figure 5A). For the small basis set sizes up to 50 basis functions, subjects' performance was now at chance level. As in Experiment 1 and 2, enlarging the basis set size resulted in an increase in the average number of correct responses. However, the relationship between basis set size and percent correct responses changed from an asymptotic to an almost linear dependency, resulting in a much smaller gain in performance when increasing the basis set size. Since at a basis set size of 250 subjects could do the task with an accuracy of about 70%, this effect is not attributable to a general incapability of the observers to do the task with the white noise patches.

The effect of changing the reconstructed data set from natural images to white noise patches is further illustrated by computing the difference between the mean percentages of correct responses in Experiment 1 and 3. Plotting this difference as a function of basis set size showed that at intermediate basis set sizes, all algorithms were least capable of representing white noise patches (Figure 5B). Overall, NMF was most affected by the image manipulation, while PCA and SPN appeared equally robust. The impact on NMF was large enough that in this experiment, NMF had the worst reconstruction performance at almost every basis set size. In contrast, its reconstruction performance had been intermediate in Experiment 1. Paired t-tests between the algorithms, performed separately at each basis set size, showed significant differences between NMF and PCA for 25 (t[7]=5.158, p=.001) and 175 BF (t[7]=6.677, p<.000). The differences to SPN became significant at a basis set size of 100 (t[7]=-2.760, p=.028) and 250 (t[7]=4.950, p=.002). The only differences between PCA and SPN occurred for 175 BF (t[7]=2.846, p=.025).

Interestingly, Experiment 3 again demonstrated a close correspondence between RM and psychophysical performance (Figures 5C and D). Here, we also found an almost linear relationship between basis set size and RM. Computing the differences between the RM in Experiments 1 and 3 in addition showed NMF to be worst in representing the white noise patches, followed by PCA and SPN.

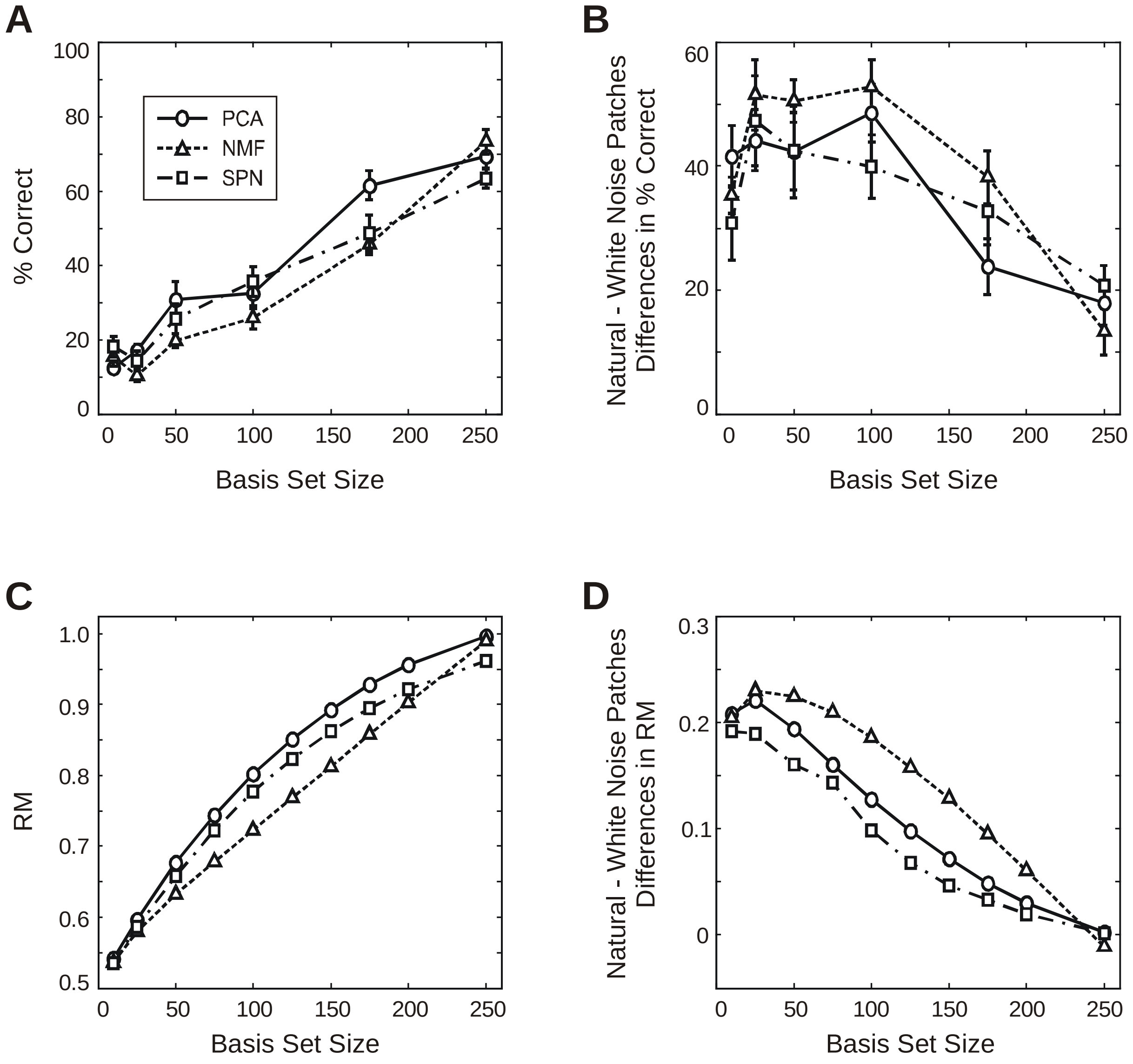

Figure 5 - Results of Experiment 3: (A,B) Psychophysical results: **(A)** Percentage of correct responses per algorithm and basis set size, averaged over all subjects. Error bars show the SEM. **(B)** Differences in percent correct responses between Experiment 1 and 3, again averaged over all subjects. Error bars depict the SEM. **(C,D)** Computational results: **(C)** RM as a function of algorithm and basis set size. **(D)** Difference in reconstruction match between Experiment 1 and 3.

6. Discussion

We have tested how well human subjects can identify natural image patches synthesized using three different algorithms (PCA, NMF and SPN). We tested algorithm efficiency by varying the number of BF that each algorithm could use to represent a large set of image patches. Not surprisingly, we found that each algorithm performed better the more BF it was allowed. Analysis of the limiting basis set size revealed that PCA performance rose most rapidly with increasing basis set size, reaching a criterion performance level of 80% correct with only 110 BF. It also reached the highest overall performance levels of 87.3% correct for 250 BF. SPN differed from PCA in that it minimizes an objective function that contains an additional term, designed to produce sparse responses for the set of BF. The introduction of this sparsity term comes at a cost – SPN performance is lower than performance using PCA as evidenced by a higher number of BF necessary to reach the criterion performance (160 BF), lower overall performance of 83.9% correct as well as statistically significant differences in performance at intermediate basis function sizes. Interestingly, BF computed using NMF led to performance that was statistically indistinguishable from performance using PCA in the present study. This is surprising, particularly since BF computed with NMF were very different from PCA BF.

In general, we find that human perceptual performance can be well predicted by an objective measure of image similarity – the mean squared reconstruction error (MSE). PCA is a method that actually minimizes the MSE, so it is not surprising that it leads to the lowest MSE values. Since SPN minimized the MSE plus a sparsity term, it generally performs worse than PCA, while NMF performance levels are intermediate between the two other methods. This pattern exactly mirrors the human psychophysical data described above. This is important, because it highlights that in the present study the MSE is a rather good predictor of human performance, at least for the image sets employed in the present study that had similar mean intensity, Fourier amplitude spectra and retinal position and size. It is likely that controlling each of these factors is quite important. For example, it is known that human ability to recognize images changes systematically with variations in the Fourier amplitude spectrum (Parraga et al 2000), so failure to control this factor would likely introduce additional variance into the results. Another recent study has demonstrated that human recognition performance is largely invariant to changes in image size (Furmanski & Engel 2000), whereas the reconstruction error (MSE) computed as in the present study would be highly susceptible to changes in size. Simple size changes would lead to a breakdown of the correspondence between MSE and perceptual performance.

By testing generalization performance of the algorithms, we evaluated whether BF were specific for the particular natural image patches that were used to compute them, or whether they were general in that they could equally represent arbitrary image patches. We found that BF were quite general, and that differences were to small to be resolved in human psychophysical performance – human subjects performed similarly for arbitrary image patches (generalization condition) and those image patches used in the BF' computation (direct condition). Examination of the MSE revealed however that with increasing basis set size, each algorithm initially achieved good reconstruction performance by representing image structure specific to the training set, and after an inflexion point began to extract more general structure. This happened at about basis set size 100 for PCA, and near 175 for both NMF and SPN. This indicates that the three algorithms initially represent the structure in their training set, but later they become more general and are able to capture arbitrary images with equal fidelity. For PCA, the set size (256 BF) always leads to an MSE of exactly zero. This becomes clear if BF and images are considered as vectors, in this case having 256 dimensions. The PCA BF are orthogonal vectors. A set of 256 BF therefore represents a full basis for a 256 dimensional space, and describing images in terms of the PCA BF is a pure transformation of the images from one space to another without a reduction in dimensionality. The images generated by the linear combination of BF at complete set size are thus identical to the original image.

The results of Experiment 3 demonstrated the effects described so far to be specific for natural image patches. Applying the algorithms to non-natural, white noise image patches resulted in a qualitatively different behavior of reconstruction performance. For all three algorithms, performance tended to show a linear dependence on basis set size. Interestingly, NMF was most affected by the change in the data set from which basis functions were computed. This is consistent with the idea that NMF represents a more biologically plausible way of decomposing visual images into BF. NMF's restriction to allow only positive contributions from each basis function or feature results in BF that are localized to particular regions of space resembling the receptive field structure in early or intermediate visual areas. The present results suggest that using NMF BF, psychophysical performance similar to the best algorithm we tested (PCA) can be supported. NMF performance does not exceed the performance of PCA, providing no support for a privileged role of NMF as an approximation of visual system performance. The human visual system is a hierarchical system containing many areas with distinct feature selectivity and with strongly non-linear tuning characteristics especially in higher cortical areas. Such complexity may need to be captured by computational models before performance benefits can be achieved. This highlights the importance of developing more realistic models with multiple layers, and computing feature maps for each layer. Along these lines, a two-layer sparse coding model can not only learn the receptive fields of simple and complex cells, but also a topography of simple cells that resembles the topography found in V1 (Hyvarinen & Hoyer 2001). Applying additional processing steps to the responses of a purely linear network layer can also lead to nonlinear behavior comparable to nonlinear properties observed in real neurons. For example, the responses of a set of linear filters derived so as to be independent from each other, nonetheless contained nonlinear dependencies when applied to natural stimuli (Schwartz & Simoncelli 2001). By removing these nonlinear dependencies through division of the response of each filter by the weighted responses of filters in its neighborhood, this model achieved a kind of gain control. This enabled the model to account for nonlinear response properties observed in real neurons like the nonlinear changes of tuning curves at different input levels.

In principle, the feature maps derived by models as the ones used here, could be compared to neural responses in different visual processing areas of the monkey brain. This comparison has previously been performed on the level of V1 simple cells. By using different input data, networks based on the same sparse coding principle that underlies SPN and NMF have been shown to generate filters that resemble the spatiotemporal (van Hateren & van der Schaaf A. 1998), as well as binocular and chromatic filter properties (Hoyer & Hyvarinen 2000) of V1 neurons. Adding further network layers to the models and performing the same comparison between feature maps of later network stages and response properties of neurons in higher cortical areas may be a critical step to answer the question how representations are assembled from simpler parts in high-level association cortex. In addition, psychophysical studies can allow quantitative evaluation of the performance of these models, and thus provide further insight into how well they approximate functions of the human visual system.

References

1. Field DJ (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. [A]* 4(12):2379-94
2. Furmanski CS, Engel SA (2000). Perceptual Learning in object recognition: object specificity and size invariance. *Vision Res.* 40(5):473-84
3. Hoyer PO, Hyvarinen A (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network.* 11(3):191-210
4. Hyvarinen A, Hoyer PO (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Res.* 41(18):2413-23
5. Joliffe IT (1986). Principal Component Analysis. New York: Springer.
6. Lee DD, Seung HS (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788-91
7. Olshausen BA, Field DJ (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images [see comments]. *Nature* 381(6583):607-9
8. Parraga CA, Troscianko T, Tolhurst DJ (2000). The human visual system is optimised for processing the spatial information in natural visual images. *Curr. Biol.* 10(1):35-8
9. Schwartz O, Simoncelli EP (2001). Natural signal statistics and sensory gain control. *Nat Neurosci* 4(8):819-25
10. van Hateren JH, van der Schaaf A (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc R. Soc. Lond B Biol Sci* 265(1394):359-66

Appendix

Images were synthesized as a combination of BF,

$$r_j(\vec{x}) = \sum_{i=1}^{L} b_i(\vec{x}) \cdot a_{ij} \ . \quad \text{(Eq. 1)}$$

Here, r_j indicates the synthesis or reconstruction of the original image o_j. b_i denotes the BF, with i ranging from 1 to the basis set size L. The basis set size L has a value between 1 and the number of pixels in the original images. Each BF has a coefficient a_{ij} assigned to it, which is specific not only for the BF but also for the image to be synthesized. Both images and BF are a function of the spatial position $\vec{x}$, i.e. a function of pixels.

Each algorithm derives BF by optimizing a specific objective function E. For PCA, this function is based on the mean squared error between reconstructed and original images:

$$E = \frac{1}{N} \sum_j \sum_{\vec{x}} \left[o_j(\vec{x}) - r_j(\vec{x}) \right]^2 .$$

BF are adapted to minimize E. This is achieved by using the eigenvectors of the images' covariance matrix as BF (Joliffe, 1986). Additionally, the BF are sorted according to the variance of their coefficients. The coefficients are then computed as:

$$a_{ij} = \sum_{\vec{x}} o_j(\vec{x}) b_i(\vec{x}).$$

NMF employs the following objective function, which is maximized by the derived BF:

$$E = \sum_{\vec{x}} \sum_j \left[o_j(\vec{x}) \log\left(r_j(\vec{x})\right) - r_j(\vec{x}) \right].$$

BF and coefficients are derived using an iterative procedure with the following update rules:

$$b_i^{new}(\vec{x}) = b_i(\vec{x}) \cdot \sum_j \frac{o_j(\vec{x})}{r_j(\vec{x})} a_{ij} ,$$

$$a_{ij}^{new} = a_{ij} \sum_{\vec{x}} b_i(\vec{x}) \frac{o_j(\vec{x})}{r_j(\vec{x})} .$$

Subsequently, the BF are normalized:

$$b_i^{new}(\vec{x}) = \frac{b_i(\vec{x})}{\sum_i b_i(\vec{x})} .$$

These update rules lead to a maximum of the objective function E under the constraints $b_i(\vec{x}) \geq 0,\ a_{ij} \geq 0$ for all $i, j,\ \vec{x}$.

The SPN objective function consists of the mean squared error between reconstructed and original images combined with a value for the sparseness of the coefficients' distribution:

$$E = \sum_{\vec{x}} \left[o_j(\vec{x}) - r_j(\vec{x}) \right]^2 + \beta \sum_i S\left(\frac{a_{ij}}{\sigma} \right).$$

S was set to $S(a) = \log\left(1 + a^2\right)$, and $\beta = 2.2$, $\sigma = 0.316$. This objective function is minimized in two phases. First, for each image o_j, E is minimized with respect to the a_{ij}, keeping the BF fixed. To find the minimum, the coefficients evolve in the direction of the gradient

$$\frac{\partial E}{\partial a_{ij}} = \sum_{\vec{x}} \left[o_j(\vec{x}) - r_j(\vec{x})\right] b_i(\vec{x}) + \frac{\beta}{\sigma} S'\left(\frac{a_{ij}}{\sigma}\right).$$

More specifically, the implemented algorithm uses a conjugate gradient descent routine to find the minimum of *E* with respect to the coefficients. In the second phase, the BF are updated such that *E* averaged over many image presentations is minimized. This yields the update rule:

$$b_i^{new}(\vec{x}) = b_i(\vec{x}) + \frac{\eta}{N} \sum_{j=1}^{N} \left(o_j(\vec{x}) - r_j(\vec{x})\right) \cdot a_{ij} .$$

The learning rate η was set to $\eta = 0.6 \frac{1}{\log(t+1)}$, were *t* is the number of iterations performed.

The BF are normalized by setting:

$$\left(\sum_{\vec{x}} \left|b_i(\vec{x})\right|^2\right)^{new} = \left(\sum_{\vec{x}} \left|b_i(\vec{x})\right|^2\right) \cdot \left(\frac{\frac{1}{N}\sum_j a_{ij}^2}{\sigma_{goal}^2}\right)^{2\alpha} ,$$

with $\sigma_{goal}^2 = 0.1$, α=0.02.