On the learning of nonlinear visual features from natural images by optimizing response energies

Jussi T. Lindgren Aapo Hyvärinen

Abstract— **The operation of V1 simple cells in primates has been traditionally modelled with linear models resembling Gabor filters, whereas the functionality of subsequent visual cortical areas is less well understood. Here we explore the learning of mechanisms for further nonlinear processing by assuming a functional form of a product of two linear filter responses, and estimating a basis for the given visual data by optimizing for robust alternative of variance of the nonlinear model outputs. By a simple transformation of the learned model, we demonstrate that on natural images, both minimization and maximization in our setting lead to oriented, band-pass and localized linear filters whose responses are then nonlinearly combined. In minimization, the method learns to multiply the responses of two Gabor-like filters, whereas in maximization it learns to subtract the response magnitudes of two Gabor-like filters. Empirically, these learned nonlinear filters appear to function as conjunction detectors and as opponent orientation filters, respectively. We provide a preliminary explanation for our results in terms of filter energy correlations and fourth power optimization.**

I. Introduction

The study of natural image statistics (see e.g. [1], [2]) examines the relations between the statistical structure of natural images and properties of visual processing. One especially important question not easily addressable in the frameworks of psychophysics or neurophysiology concerns the functional purpose of the encountered visual machinery, i.e. "why is it like it is?". This question implies goals for the visual processing and is slowly becoming more addressable (for advocates of this viewpoint, see e.g. [3], [4]). Regarding vision, one way to approach this issue is to fit computational models to natural image data, optimizing the model to fulfill some chosen objective function. If the optimization leads to similar processing as encountered in natural systems, this gives an interesting proposition regarding the functional properties of the natural processing. On the other hand, conflicting results enable to question and refine the used objective in the studied setting. Finally, understanding the statistical structure of image data is also clearly useful for applied fields such as computer vision and content based image retrieval.

Usual linear approaches to natural image modelling – such as Independent Component Analysis (ICA) and sparse coding – lead to image models that represent images as linear combinations of simple elements such as edges and bars that resemble Gabor filters or V1 simple cell receptive fields of primates [5], [6], suggesting that such processing provides an efficient low-level coding for the visual input. However, it should be kept in mind that such simple features may be insufficient for more complex tasks, such as segmentation or object recognition. This intuition is complemented by results from physiology, where proceeding onward in the visual hierarchy from the area V1 leads to increasingly nonlinear processing. Already in V2, neurons appear to respond to feature conjunctions, but not to the individual features [7], [8], [9]. Such behavior can not be easily attained with linear models [10], and thus nonlinear processing appears a prerequisite for models hoping to have resemblance to later-stage visual functionality.

Unfortunately, there are several difficulties with nonlinear models, and questions include how to choose the model structure, how to fit its parameters properly, and how to interpret the model. Here we take a small but natural step forward from linear models by considering models consisting of products of responses of two linear filters, i.e. we consider a subclass of quadratic models that have been previously studied also in the context of natural images [11], [12], [13], [14], [15], [16]. In the current paper we give evidence that although previous studies using this model class have largely led to model behavior resembling that of complex cells, this may have more to do with the used objectives than the model class or the data. Typically, independence of the individual quadratic component outputs has been optimized ([11], [12], [13], [14], [15]) but also temporal coherence [13] and slow feature analysis [17] have been studied in this context. Here, we examine an objective related to maximizing or minimizing the energy of the component responses (in our setting this is closely related to response variances). Maximization of our objective could be suggested to correspond to searching for maximally active directions in the product space of linear filter responses, whereas minimization might correspond to looking for sparse feature combinations. Both cases are nontrivial, as in our setting the responses of the underlying linear filters that are combined have unit variance. We show that depending on the optimization direction, models having either opponent orientation behavior or conjunctive behavior are learned. We show that both types of models produce their outputs by nonlinearly combining outputs of linear filters that are localized, oriented and bandpass, resembling Gabor filters. It should be noted that qualitatively similar behavior has also been encountered in natural systems. Opponent orientation behavior has been found in second-order processes of the human vision [18] (but see also [19]), whereas conjunctive processing preferring angles and corners has been demonstrated in the macaque area V2 [7], [8], [9]. The current paper presents (to the best of our knowledge)

The authors are with the Department of Computer Science and HIIT, University of Helsinki, Finland (email: firstname.lastname@cs.helsinki.fi).

the first computational study to demonstrate the emergence of opponent orientation behavior, and the first study to show that both conjunctive and subtractive behavior can emerge from the same computational optimization framework (conjunctive feature learning has been previously observed in [15] but with a more complex approach).

The rest of this paper is organized as follows. In section II we describe our used method. Section III describes the data and our experimental setup, and the empirical results are presented in section IV. We provide some theoretical explanations for our results in section V. Finally, section VI concludes the paper.

II. Optimization of paired energies

Assume vectorized image patches $\mathbf{x} \in \Re^{2m}$, $E[\mathbf{x}] = \mathbf{0}$ and that $E[\mathbf{x}\mathbf{x}^T] = \mathbf{I}$, i.e. the data has zero mean and its covariance matrix is the identity matrix. This can be always be attained with a linear whitening transform, see e.g. [20]. For the whitened data, we attempt to learn an orthonormal basis $\mathbf{W} \in \Re^{2m \times 2m}$ for $\mathbf{x}$, with rows of $\mathbf{W}$ as linear projection directions. Instead of working in the output space of the linear filter outputs, i.e. with responses of the form $s_i = \mathbf{w}_i^T\mathbf{x}$ for filter $\mathbf{w}_i$, we assume a functional form of a product of two linear filter responses, that is, the response of a single nonlinear component i in our model is

$$s_i(\mathbf{x}) = (\mathbf{v}_i^T\mathbf{x})(\mathbf{w}_i^T\mathbf{x}) \quad (1)$$

$$= \frac{1}{4}(\mathbf{a}_i^T\mathbf{x})^2 - \frac{1}{4}(\mathbf{b}_i^T\mathbf{x})^2, \quad (2)$$

where $\mathbf{v}_i$ and $\mathbf{w}_i$ are the i:th and $m+i$:th rows of $\mathbf{W}$, respectively. The second identity follows by simple manipulation from the choice $\mathbf{a}_i = \mathbf{v}_i + \mathbf{w}_i$ and $\mathbf{b}_i = \mathbf{v}_i - \mathbf{w}_i$. This computation is illustrated in network form in Figure 1. It can thus be seen that the function class we are optimizing is a subclass of quadratic models that have been previously considered for natural images [11], [12], [13], [14], [15], and in modelling texture processing (e.g. [21]). Two-layer models in general (such as that of [22]) closely resemble quadratic models. The subclass chosen here has the benefit that it has only $O(m)$ parameters and is more constrained, whereas a full quadratic model would require $O(m^2)$ parameters. Also, once the models have been fitted, the operation of a single filter product is easier to explain than a full quadratic model, which supports a linear combination of $2m$ terms instead of the subtraction of two terms in eq. (2). That is, a full quadratic model can linearly filter the response energies of $2m$ linear filters; here this second-stage filtering is constrained to be a simple fixed subtraction of two response energies.

The actual objective function that is used to optimize $\mathbf{W}$ can strongly affect the learned models. In our setting, the objective L_i of a single component i is

$$L_i = E_{\mathbf{x}}[g(s_i(\mathbf{x}))], \quad (3)$$

where g is a nonlinearity. We estimate the weights of $\mathbf{W}$ by either minimizing or maximizing a global objective function,

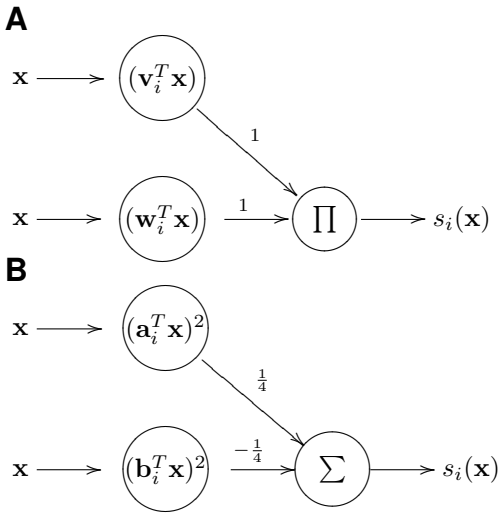

Fig. 1. Two ways of representing our nonlinear components as simple networks. A) A single component s_i pairing two filter outputs $\mathbf{v}^T\mathbf{x}$ and $\mathbf{w}^T\mathbf{x}$ by multiplying them. B) An equivalent form that takes the difference of the energies of two filters $\mathbf{a}$ and $\mathbf{b}$ that follow from $\mathbf{v}$ and $\mathbf{w}$ by simple manipulation (see text).

which is simply a sum over all of the individual m objectives, giving

$$L = \sum_{i=1}^{m} L_i = \sum_{i=1}^{m} E_{\mathbf{x}}[g(s_i(\mathbf{x}))]. \quad (4)$$

In similar spirit to previous work on linear models, we could use g for various purposes, such as preferring that the response s follows some chosen distribution (as in e.g. [11]), or is as sparse as possible (e.g. [5]), or that the different filter responses are maximally independent by maximization of nongaussianity (e.g. [20]). For example, if we had specified the more traditional summation of the two terms in eq. (2) instead of subtraction, we could expect to learn complex cell models if we chose g to prefer sparse outputs [11]. On the other hand, by specifying a subtraction (or equivalently the multiplication of two linear filter outputs) we are potentially able to get more tightly tuned features that require high responses from both paired linear filters for the magnitude of the product response to be high (as in [15]).

Instead of the above choices, in the current paper we study three measures of response energy,

$$g_1(s) = \mathrm{abs}(s), \quad (5)$$

$$g_2(s) = s^2, \quad (6)$$

$$g_{lc}(s) = \log\cosh(s), \quad (7)$$

where g_1 and g_{lc} can be taken as more robust versions of the (unsigned) energy g_2. As $\log\cosh$ grows substantially slower than s^2 after $|s| > 1$, it is less subjective to effects of outliers than s^2, whereas $\mathrm{abs}()$ naturally has linear behavior also in region $[-1, 1]$. These simple objective functions are shown in Figure 2 for clarity; in section V we show how s^2 is related to both the optimization of energy correlation of linear filters and the fourth powers of the linear filter responses.

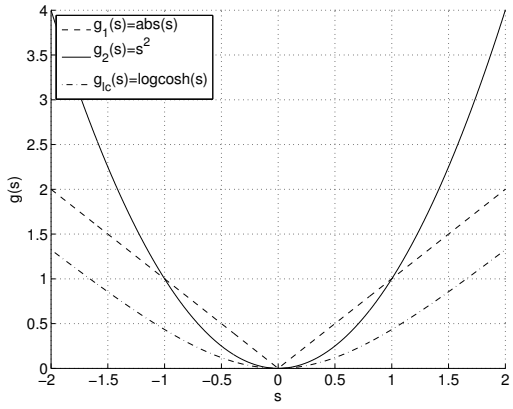

Fig. 2. Behavior of some possible nonlinearities g. Inside the region $[-1, 1]$, $\log\cosh$ behaves similarly to s^2 and outside it approximately linearly. In the current paper the two other functions are used as more robust alternatives for s^2.

We optimize our objective functions by gradient ascent or descent, depending on if we are minimizing or maximizing. The derivatives of our nonlinearities g are

$$g_1'(s) = \text{sign}(s) \quad (8)$$
$$g_2'(s) = 2s \quad (9)$$
$$g_{lc}'(s) = \tanh(s), \quad (10)$$

and the gradients simply

$$\frac{\delta L_i}{\delta \mathbf{v}_i} = E_{\mathbf{x}}[g'(s_i)(\mathbf{w}_i^T\mathbf{x})\mathbf{x}] \quad (11)$$
$$\frac{\delta L_i}{\delta \mathbf{w}_i} = E_{\mathbf{x}}[g'(s_i)(\mathbf{v}_i^T\mathbf{x})\mathbf{x}]. \quad (12)$$

This gives the gradient ascent update rules

$$\mathbf{v}_i = \mathbf{v}_i + E_{\mathbf{x}}[g'(s_i)(\mathbf{w}_i^T\mathbf{x})\mathbf{x}] \quad (13)$$
$$\mathbf{w}_i = \mathbf{w}_i + E_{\mathbf{x}}[g'(s_i)(\mathbf{v}_i^T\mathbf{x})\mathbf{x}] \quad (14)$$

for all components $i \in 1, ..., m$. The version for minimization is simply obtained by subtracting the gradients instead of adding them. After each gradient update, we project $\mathbf{W}$ to the constraint set (orthonormality). This projection is

$$\mathbf{W} = ((\mathbf{W}\mathbf{W}^T)^{-1/2})\mathbf{W}, \quad (15)$$

for details see [23], [20].

Finally, note that although we estimate parameters for nonlinear functions $s(\mathbf{x})$, our basis $\mathbf{W}$ was still defined as orthonormal. This means that $\mathbf{W}$ is invertible, and that no information is lost in the linear filtering stage, regardless of the parameters the learning chooses.

The whole algorithm is summarized as pseudocode in Figure 3. The matlab sources of the implementation will be made publicly available, accessible from the first authors homepage.

Whiten the data, possibly after reducing dimension;
Pick random, orthonormal initial $\mathbf{W}$;
% Perform gradient ascent
while $\neg$ *converged* **do**
  for $i \in 1, ..., m$ **do**
    $s \leftarrow (\mathbf{v}_i^T\mathbf{x})(\mathbf{w}_i^T\mathbf{x})$;
    $\mathbf{v}_i \leftarrow \mathbf{v}_i + E_{\mathbf{x}}[g'(s)(\mathbf{w}_i^T\mathbf{x})\mathbf{x}]$;
    $\mathbf{w}_i \leftarrow \mathbf{w}_i + E_{\mathbf{x}}[g'(s)(\mathbf{v}_i^T\mathbf{x})\mathbf{x}]$;
  end
  % Orthonormalize
  $\mathbf{W} \leftarrow \mathbb{R}e((\mathbf{W}\mathbf{W}^T)^{-1/2})\mathbf{W}$;
end
Transform $\mathbf{W}$ to the original space;

Fig. 3. The used gradient ascent algorithm for maximization of the used objective function as pseudocode. Minimizing version is obtained simply by flipping the signs of the gradient updates.

III. Experimental setup

In our experiments we used the natural image dataset provided by van Hateren and van der Schaaf [6]. This dataset contains over $4,000$ grayscale images representing natural scenes, each image having a size of 1024×1536. We used the '.iml' versions of the images, and cropped 4 pixels from all sides of the images to avoid border anomalies present in the data. Then, we performed 3×3 block averaging on each image to reduce the effect of noise and sampling artifacts. Finally, we applied a natural logarithm to each image to balance the very long right tail of the pixel intensity distribution, effectively compressing the dynamic range of the images. Both the averaging and the logarithm are commonly used to address difficulties arising from using raw natural images in computational modelling (e.g. [5], [6], [24]). It is also well-known that retinal processing performs similar averaging and compressive transform (see e.g. the references in [6]).

After preprocessing each image as a whole, we sampled a training set of $200,000$ small patches from the images, each patch having a resolution of 16×16 pixels. We then subtracted the local DC-component (mean intensity) from each patch and removed the mean of the entire dataset (i.e. each variable had zero mean). These patches then formed the 256-dimensional data we used to optimize our models. We also prepared another set of $20,000$ patches similarly to use as a test set. All objective functions and test measurements related to the data in the section IV were computed using this test set to avoid the reported results being due to overfitting.

Before applying the gradient method, the training data was whitened by a linear transform through Principal Components Analysis (PCA, see e.g. [20]), while retaining $200/256$ of the most significant principal components for estimation of 100 filter pairs. Note that whitening done in this way is a standard procedure e.g. in Independent Component Analysis in general [20] and similar processing is commonly used also in low level visual modelling (e.g. [5], [11], [15]), and as with the averaging and the compressive transform, whitening also appears to have its counterpart in natural early visual processing [25]. After learning the models in the whitened

space, we transformed them back to the original space for visualization and analysis.

IV. Results

Of the three objective functions we studied, the function $g_1 = s^2$ appeared to require more examples than the other two to reach similar behavior of the learned models, and yet upon convergence the filter masks did not look as structured as they did with the other choices. This is possibly due to the lack of robustness in s^2. Function $g_{lc} = \log\cosh(s)$ behaved well in maximization and produced very similar results to $g_1 = \mathrm{abs}(s)$, but the results of $\log\cosh$ did not appear as good in the minimization setting. This may be connected to $\log\cosh$ behaving like s^2 near the origin. In both maximization and minimization, abs performed well by producing visually clear filters, and in the following account of our empirical results, we only report those related to g_1.

It turned out that in our setting, both the maximization and minimization of the objective function of eq. (4) create meaningful but dual results. Figure 4 illustrates 64 of the 100 learned filter pairs obtained from the maximization of the objective function. In each quadruple, the top two filters correspond to filters $\mathbf{w}$ and $\mathbf{v}$, and the bottom row filters to $\mathbf{a}$ and $\mathbf{b}$. Either of these two pairs is sufficient to compute the pair response according to eqs. (1) and (2). The actually learned basis $\mathbf{W}$ would consists of the top row filters. The 100 quadruples were sorted according to the increasing objective function $E_{\mathbf{x}}[|s|]$ as measured on the test set, after which 64 components were selected for display by skipping a little fewer than every second component in the sorted order. The omitted components were similar to those shown.

Looking at Figure 4 shows that in the maximization case, the features $\mathbf{a}$ and $\mathbf{b}$ are oriented, localized and bandpass, resembling Gabor- or Haar-filters, not unlike those found in numerous previous studies (e.g. [5], [6], [11], [17]). However, filters $\mathbf{w}$ and $\mathbf{v}$ that were explicitly fitted are mixtures of these properties. This situation gets reversed in the minimization case (see Figure 5), where now filters $\mathbf{w}$ and $\mathbf{v}$ have the familiar properties, whereas $\mathbf{a}$ and $\mathbf{b}$ are their mixtures.

If these learned nonlinear filters are used on images to compute the responses $s_i(\mathbf{x})$ in a convolutive fashion, the filters from the maximization case appear to behave as opponent orientation filters due to $(\mathbf{b}^T\mathbf{x})^2$ getting subtracted from $(\mathbf{a}^T\mathbf{x})^2$, i.e. the component compares the response energy of $\mathbf{b}$ against that of $\mathbf{a}$, as can be seen from eq. (2). Figure 6A shows a simple test image, and Figure 6B the response behavior for the component that had the best objective value in the maximization setting. In the minimization setting the learned filters are instead highly specific and react strongest when both responses $\mathbf{w}^T\mathbf{x}$ and $\mathbf{v}^T\mathbf{x}$ are high in magnitude, as is apparent from eq. (1). This behavior can be seen in Figure 6C, showing that the pair responds strongly only to a right angle in a preferred orientation. Hence although in both cases the participating filters in the pair may be arranged similarly, there is a strong difference in the component operation depending on the form where the filters appear as Gabors: if its that of eq. (2), we have subtractive filters, and in that of eq. (1), conjunctive.

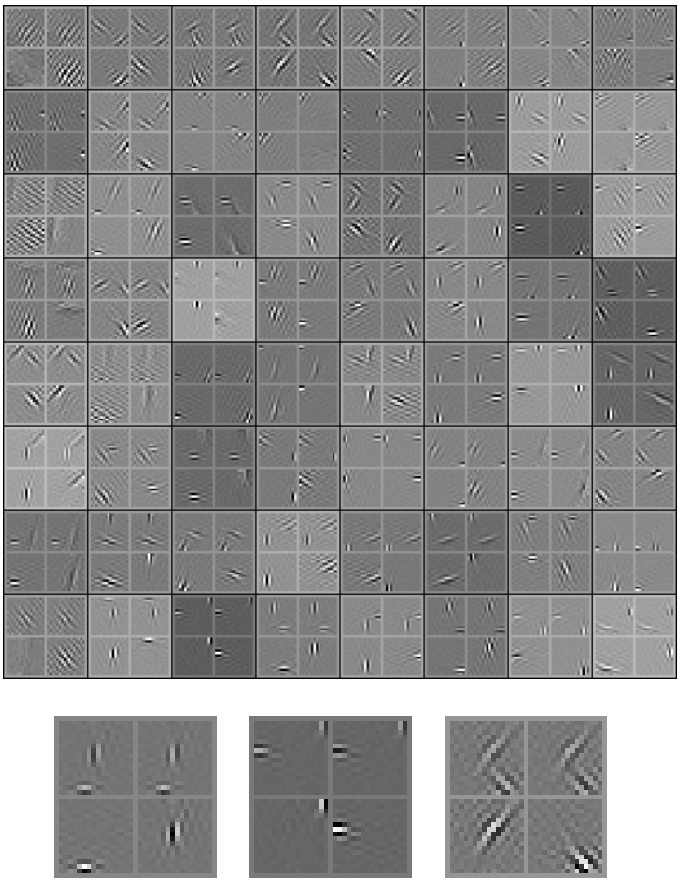

Fig. 4. Top, the $64/100$ components learned in the maximization setting for g_1 nonlinearity. In each quadruple, the two upper filters are $\mathbf{w}$ and $\mathbf{v}$, and the lower filters $\mathbf{a}$ and $\mathbf{b}$. Either pair is sufficient to compute the response s. Notice that in this case, the vectors $\mathbf{a}$ and $\mathbf{b}$ resemble Gabor-filters. The quadruples have been sorted according to the objective value on test set, growing from left to right, top to bottom, indicating that the best pair in terms of the objective is at the bottom right. Each quadruple has also been scaled to $[0, 1]$ range separately and the whole image has been contrast enhanced for printing. The individual filters do not have a DC component, the effect is due to the scaling. Bottom, close-up of three of the filters.

An important question regards whether the learned pairings reflect structure in the input data or not (which was left an open question for conjunctive features learned with another approach in [15]). The fact that the ordering of the components by their objective values shows some structural changes in the used pairings in figures 4 and 5 already suggests that the pairings are non-arbitrary. Especially filters arranged in right angles are preferably paired by both maximization and minimization, giving high and low objective values, respectively. To further illustrate this structure, we show histograms of the orientation differences of the paired filters in Figure 7. We considered the component filters in their Gabor-like forms and measured each filter angle simply by finding the maximum of each filter masks Fourier power spectrum and then obtaining the angle from polar coordinate transform of the maximum location. As can be seen, generally the angle differences between the filters

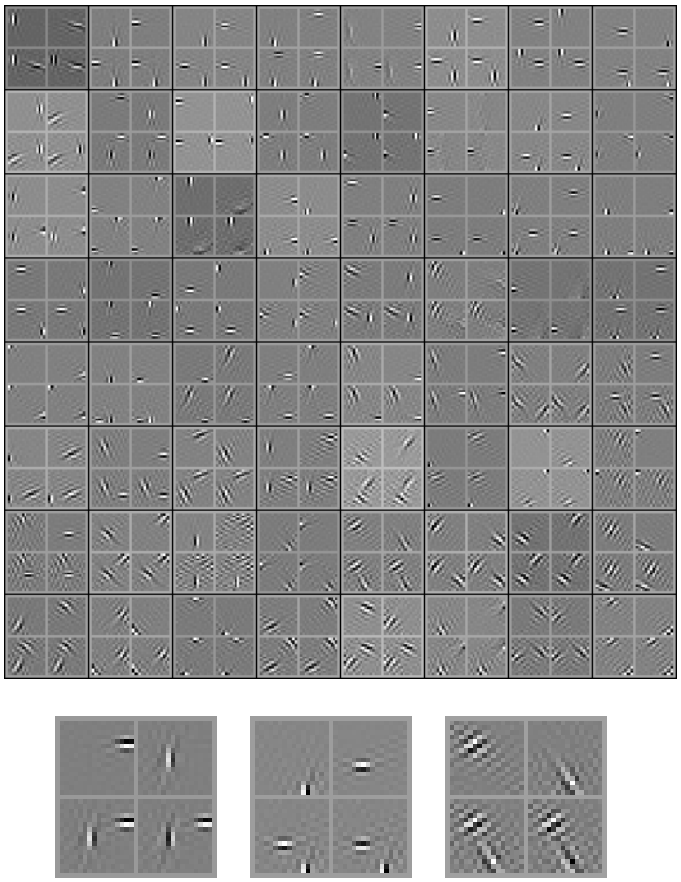

Fig. 5. Top, the basis vectors learned in the minimization setting for g_1 nonlinearity. The figure is read as Figure 4. Now vectors $\mathbf{w}$ and $\mathbf{v}$ resemble Gabor-filters and since the task was minimization, the worst pair in terms of minimization is at the bottom right. Bottom, close-ups of three of the filters.

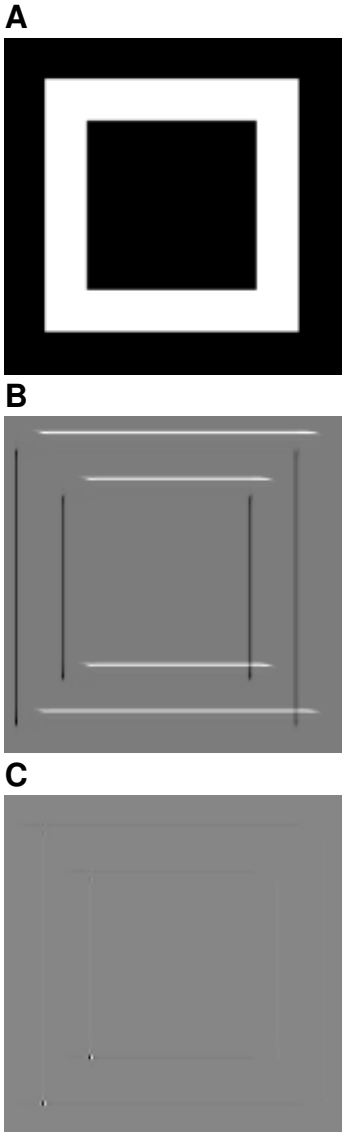

Fig. 6. Responses from filtering a test image with two of the learned filters. A) The original image. B) An image resulting from filtering with the filter having the best objective value in the maximization setting. This filter reacts positively to horizontal edges and negatively to vertical edges. Hence the filter implements opponent orientation behavior. C) An image resulting from the best filter from the minimization framework. Now the filter reacts highly only to co-occurrence of horizontal and vertical edges as positioned in the lower left corner of the test image in A. The filter operation thus resembles a corner detector. Note that both of the filters appear phase invariant.

in both minimization and maximization setting appear to be concentrated on right angles ($\pm 90^\circ$ or $\pm\pi/2$ radians). Curiously, similar angle bias has also been noted in the macaque V2 [9] although there the majority of the cells found preferred collinear receptive fields. As seen from Figure 7, collinear filters appear to be totally absent from the results of our current method, whereas collinear combinations have been previously observed to emerge from a full quadratic model setting with a different objective [15]. We are currently studying the reason for this difference.

We also examined the possible arbitrariness of the pairings by estimating what kind of global objective values could be attained by simply making sets of random pairs out of the learned filters (with the filters in their Gabor forms). We selected all the filters learned by our method in either maximization or minimization setting, and used as a pairing-constraint that each filter can appear only in a single pair. We evaluated $100,000$ sets of random pairs against the method-chosen pair set, and a pairing set made by greedy selection. The greedy pairing was done by simply selecting the best filter pair in terms of the objective function (on training data) until no filters were left. Then the global objective value (as in eq.(4)) of each set was evaluated on the patches of the test set. The histograms in Figure 8 show that the global objective function values for the sampled sets commonly are worse than those reached by either the gradient method or by the greedy pairing, which have been marked onto the histograms as dotted and continuous lines, respectively.

A simple hypothesis regarding our results is to suggest that the method simply tries to keep the paired filters as far away spatially from each other as possible, in the form where they resemble Gabors (due to dependencies in natural images typically decreasing with distance, see e.g. [26]). This doesn't seem to be trivially the case (or the gradient method does not succeed in this). We examined this question by

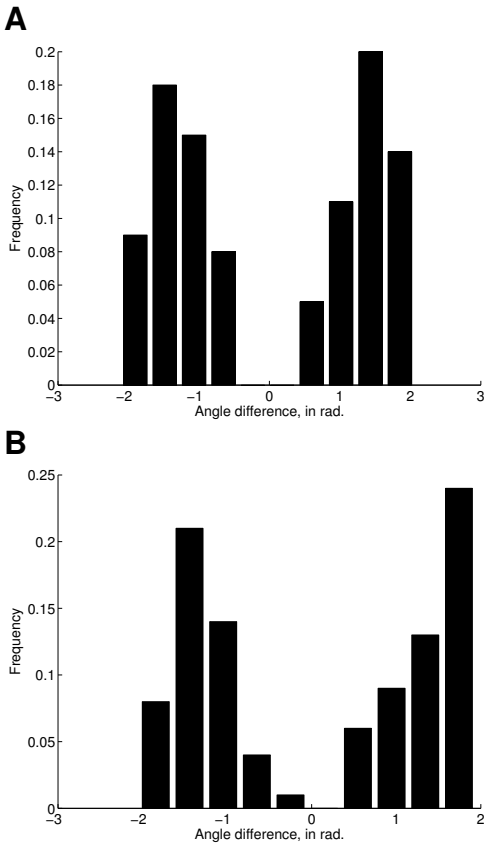

Fig. 7. Differences in orientations of the learned filter pairs when they have been transformed into their Gabor-like form. A) Maximization case, pairing filters $\mathbf{a}$ and $\mathbf{b}$. B) Minimization case, pairing filters $\mathbf{w}$ and $\mathbf{v}$.

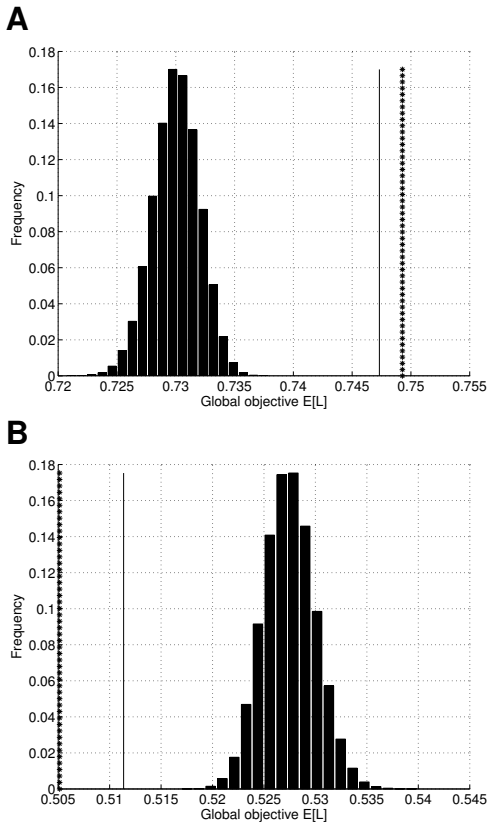

Fig. 8. Histograms of average objective function values for a sample of $100,000$ sets of pairs made by arbitrarily pairing the filters learned by the method. A dotted line marks the average objective function value for the solution returned by the gradient method, with the continuous line giving the objective for a set made by greedy pairing of the learned filters. In both cases the pairing returned by the gradient method reaches objective function values clearly better than those commonly obtainable by random pairing. A) Maximization case, pairing filters $\mathbf{a}$ and $\mathbf{b}$. B) Minimization case, pairing filters $\mathbf{w}$ and $\mathbf{v}$.

computing the spatial Euclidean distances between the mass centroids of the paired filters, and compared this distance to the objective value of the component. This is illustrated in Figure 9, and no clear dependency can be seen. The correlation coefficient c between the two quantities was $c = 0.07$ for the maximization case ($p > 0.50$) and $c = -0.09$ for minimization ($p > 0.35$), i.e. simple correlation between mask distance and objective function value does not appear to be supported by the data for the estimated filters. This suggests that the method takes into account more complex considerations than the simple spatial distances between the paired filters.

Finally, we examined how close the objective of using g_1 is to that of g_2 in our setting. This is shown in Figure 10, where we computed $E[g_2(s)] = E[s^2]$ for each learned pair on the test set, and then plotted it against the objective function $E[g_1(s)] = E[\mathrm{abs}(s)]$ with the same data. A clear relationship can be seen, with the actual correlation coefficient c between the two quantities being $c = 0.61$ for the maximization case and $c = 0.76$ for minimization, with $p < 0.01$ for each. This shows some supportive evidence for our use of g_1 as a robust alternative to g_2.

V. Discussion

Learning quadratic models has typically led to components that have simple or complex cell properties [11], [12], [13], [14], [17]. Our results, together with those in [15], demonstrate that very different but still seemingly reasonable processing can be obtained by small alterations on the objectives and constraints. Interestingly, both subtractive ("orientation opponency") processing [18] and conjunctive ("corner detectors") processing [7], [18], [8], [9] have been noted in stages after V1, although the functional significance of such processing appears an open question.

In the current context it seems possible to ask why the models learned in our setting end up expressing such behavior. Especially, why are the filters paired as they are, and why do the models have an aspect of Gabor filters in both minimization and maximization case, depending on the representation of the model? Here we provide a preliminary sketch that connects the used optimization to the objective functions for the responses of the underlying linear filters.

First, we show how our objective function is connected to the energy correlation of the paired linear filters. Assume that we maximize the individual objective of eq. (3) for $g(s) = s^2$, since this is technically the most tractable choice. Starting from the definition of energy correlation for responses $\mathbf{v}^T\mathbf{x}$

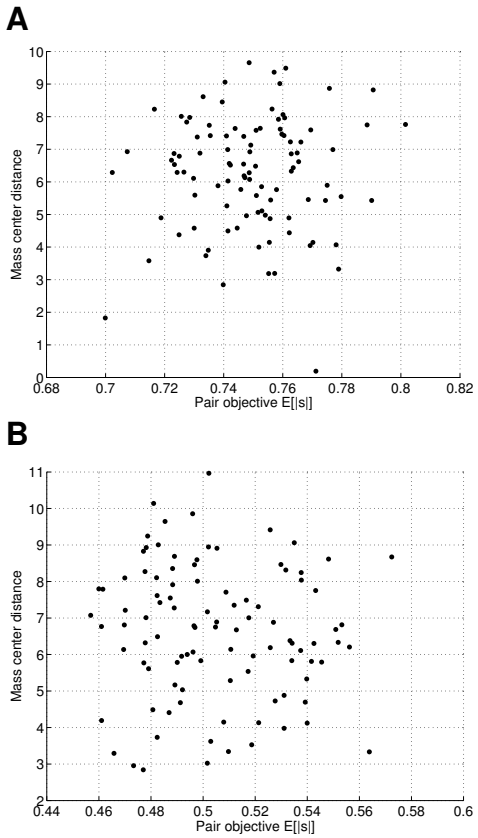

Fig. 9. Plots of the objective function value of component responses against the Euclidean distances between the spatial mass centers of the filters paired by the component. A) Maximization, correlation coefficient $c = 0.07, p > 0.50$. B) Minimization, $c = -0.09, p > 0.35$. A clear relationship is not apparent in either case.

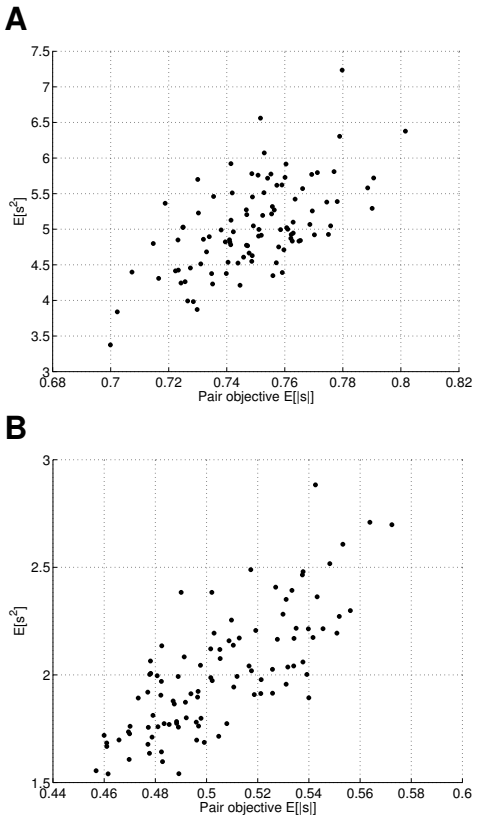

Fig. 10. Plots of the objective function value of individual components $E[\mathrm{abs}(s)]$ against the energies $E[s^2]$ of the responses. A) Maximization, correlation coefficient $c = 0.61, p < 0.01$. B) Minimization, correlation coefficient $c = 0.76, p < 0.01$.

and $\mathbf{w}^T\mathbf{x}$, we get

$$\begin{aligned}
& \mathbf{cov}_{\mathbf{x}}[(\mathbf{v}^T\mathbf{x})^2, (\mathbf{w}^T\mathbf{x})^2] && (16)\\
= \; & E_{\mathbf{x}}[((\mathbf{v}^T\mathbf{x})^2 - E[(\mathbf{v}^T\mathbf{x})^2]) \\
& \quad ((\mathbf{w}^T\mathbf{x})^2 - E[(\mathbf{w}^T\mathbf{x})^2])] && (17)\\
= \; & E_{\mathbf{x}}[((\mathbf{v}^T\mathbf{x})^2 - 1)((\mathbf{w}^T\mathbf{x})^2 - 1) && (18)\\
= \; & E_{\mathbf{x}}[(\mathbf{v}^T\mathbf{x})^2(\mathbf{w}^T\mathbf{x})^2 \\
& \quad -(\mathbf{v}^T\mathbf{x})^2 - (\mathbf{w}^T\mathbf{x})^2 + 1] && (19)\\
= \; & E_{\mathbf{x}}[(\mathbf{v}^T\mathbf{x})^2(\mathbf{w}^T\mathbf{x})^2] - 1 && (20)\\
= \; & E_{\mathbf{x}}[g(s)] - 1, && (21)
\end{aligned}$$

i.e. our objective function for a single component with an unimportant additive constant. Above, we used the linearity of expectation and $E[(\mathbf{w}^T\mathbf{x})^2] = 1$ and $E[\mathbf{w}^T\mathbf{x}] = 0$ for all $\mathbf{w}$. These two latter properties are easy to show and are due to the whitening of the data [20]. Thus, optimization in the studied setting using g_2 nonlinearity equals optimizing filter energy correlations of the paired filters.

Next, we provide a preliminary explanation why Gabor-like filters emerge in both cases. Notice that we can turn the representation of our model in eq. (1) to that of eq. (2) by the choice of $\mathbf{a} = \mathbf{v}+\mathbf{w}$ and $\mathbf{b} = \mathbf{v}-\mathbf{w}$, equal to $\mathbf{v} = 1/2(\mathbf{a}+\mathbf{b})$ and $\mathbf{w} = 1/2(\mathbf{a} - \mathbf{b})$. Now we can represent our objective function with $\mathbf{a}$ and $\mathbf{b}$ as

$$\begin{aligned}
& E_{\mathbf{x}}[(\mathbf{v}^T\mathbf{x})^2(\mathbf{w}^T\mathbf{x})^2] && (22)\\
= \; & E_{\mathbf{x}}[(\frac{1}{2}(\mathbf{a}+\mathbf{b})^T\mathbf{x})^2(\frac{1}{2}(\mathbf{a}-\mathbf{b})^T\mathbf{x})^2] && (23)\\
= \; & \frac{1}{16}E_{\mathbf{x}}[(\mathbf{a}^T\mathbf{x})^4 + (\mathbf{b}^T\mathbf{x})^4 - 2(\mathbf{a}^T\mathbf{x})^2(\mathbf{b}^T\mathbf{x})^2]. && (24)
\end{aligned}$$

Formulating our objective function this way now makes clear that in the maximization case, we actually wish to maximize terms closely resembling kurtosis (or the fourth moment) for the responses of filters $\mathbf{a}$ and $\mathbf{b}$, balanced with minimizing their energy correlation. Further, by substituting $\mathbf{a} = \mathbf{v} + \mathbf{w}$ and $\mathbf{b} = \mathbf{v} - \mathbf{w}$ to the last term on the right of eq. (24), and manipulating, we get

$$\begin{aligned}
& E_{\mathbf{x}}[(\mathbf{v}^T\mathbf{x})^2(\mathbf{w}^T\mathbf{x})^2] && (25)\\
= \; & \frac{1}{12}E_{\mathbf{x}}[(\mathbf{a}^T\mathbf{x})^4 + (\mathbf{b}^T\mathbf{x})^4] \\
& -\frac{1}{6}E_{\mathbf{x}}[(\mathbf{v}^T\mathbf{x})^4 + (\mathbf{w}^T\mathbf{x})^4]. && (26)
\end{aligned}$$

From the above expression it can be seen that in the minimization case, it benefits the optimization to minimize kurtosis for the representation using $\mathbf{a}$ and $\mathbf{b}$, while maximizing it for the representation using $\mathbf{v}$ and $\mathbf{w}$, with twice the weight. In the maximization case, these roles are reversed. This surprising dualism-like property between minimization and

maximization in our setting suggests why both optimization directions learn Gabor-like filters from natural images: this is due to the objective preferring fourth moment to be maximized in both cases but for different representations of the model. Such fourth moments for whitened data are intimately connected to kurtosis, one possible objective to use in Independent Component Analysis (see e.g. [20]), a method that is known to produce Gabor-like filters on natural images [6]. Clearly the shape of the function $f(x) = x^4$ also suggests that the objective prefers heavy-tailed (or sparse) distributions from its inputs, giving a connection to sparse coding. However, a question still remains why the kurtosis minimization in eq. (26) does not cancel the maximization out and produce very different overall results. Our results indicate that this could be due to the structure of natural images and the dependencies between the two model representations: using projections with high kurtosis appropriately in eq. (26) allows for better objective function values in both cases.

Our results allow us to speculate on an interesting possibility that the low-level processing elements in natural systems resembling e.g. Gabor filters may be by-products of higher-level non-linear mechanisms, and not necessarily optimized for some low-level purpose per se. Subsequently these mechanisms may be performing optimally for some different purpose than the one perceived from isolated studies of the lower-level machinery. Our results present an interesting contrast to the previous studies, as in our setting no explicit sparseness of the response was optimized; on the contrary, especially the maximization of energy could be seen as a kind of anti-sparseness. Yet, due to the interaction of the used function class and the objective function, kurtosis (or sparseness) still appeared to emerge as a property that should be optimized for the underlying linear filters.

We are currently working on formulating similar analytical understanding for nonlinearities $g_1 = \mathrm{abs}(s)$ and $g_{lc} = \log\cosh(s)$ as we demonstrated here for $g_2 = s^2$. It remains a possibility that the other two nonlinearities pose such objectives for the underlying linear filter responses that are different in some significant manner from those following from s^2.

VI. Conclusion

We have empirically shown that maximization or minimization of output energy in a subclass of quadratic models can lead to emergence of nonlinear filters expressing either subtractive or conjunctive behavior, respectively. We then analytically demonstrated that optimization of the square objective function for the used model class can be explained as optimization of filter energy correlations for paired linear filters, but also that this objective has intimate connections to optimization of individual fourth powers of each filter response. Our results suggest a possibility that observed conjunctive and subtractive processing in natural systems or learned computational models may have a connection to optimization of energy correlations for functions from a suitable nonlinear model class.

Acknowledgments: This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

[1] E. P. Simoncelli. Statistical modeling of photographic images. In A. Bovik, editor, *Handbook of Image and Video Processing, 2nd edition*, pages 431–441. Academic Press, 2005.

[2] G. Felsen and Y. Dan. A natural approach to studying vision. *Nature Neuroscience*, 8(12), 2005.

[3] D. Marr. *Vision*. Freeman, 1982.

[4] P. W. Glimcher. *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics*. The MIT Press, 2003.

[5] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[6] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc.R.Soc.Lond. B*, 265:359–366, 1998.

[7] J. Hegdé and D. C. van Essen. Selectivity for complex shapes in primate visual area V2. *The Journal of Neuroscience*, 20(5):RC61–66, 2000.

[8] M. Ito and H. Komatsu. Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J. Neu.sci.*, 24(13):3313–3324, 2004.

[9] A. Anzai, X. Peng, and D. C. van Essen. Neurons in monkey visual area V2 encode combinations of orientations. *Nature Neuroscience*, 10(10):1313–1321, 2007.

[10] G. Krieger and C. Zetzsche. Nonlinear image operators for the evaluation of local intrisic dimensionality. *IEEE Transactions on Image Processing*, 5(6):1026–1042, 1996.

[11] A. Hyvärinen and P. O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[12] H. Bartsch and K. Obermayer. Second-order statistics of natural images. *Neurocomputing*, 52-54:467–472, 2003.

[13] W. Hashimoto. Quadratic forms in natural images. *Network: Computation in Neural Systems*, 14(4):765–788, 2003.

[14] F. Theis and W. Nakamura. Quadratic independent component analysis. *IEICE Trans. Fundamentals*, E87-A(9):2355–2363, 2004.

[15] J. T. Lindgren and A. Hyvärinen. Emergence of conjunctive visual features by quadratic independent component analysis. In *Advances in Neural Information Processing Systems 19 (NIPS)*, pages 897–904, 2007.

[16] U. Köster and A. Hyvärinen. A two-layer ICA-like model estimated by score matching. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN2007)*, pages 798–807, 2007.

[17] P. Berkes and L. Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, 5:579–602, 2005.

[18] I. Motoyoshi I. and F. A. Kingdom. Orientation opponency in human vision revealed by energy-frequency analysis. *Vision Research*, 43(9):2197–2205, 2003.

[19] N. Graham and S. S. Wolfson. Is there opponent-orientation coding in the second-order channels of pattern vision? *Vision Research*, 44(27):3145–3175, 2004.

[20] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

[21] N. Prins and F. A. Kingdom. Direct evidence for the existence of energy-based texture mechanisms. *Perception*, 35(8):1035–1046, 2006.

[22] A. P. Johnson and C. L. Baker. First- and second-order information in natural images: a filter-based approach to image statistics. *Journal of the Optical Society of America A*, 21(6), 2004.

[23] D. Luenberger. *Optimization by Vector Space Methods*. Wiley, 1969.

[24] B. Willmore, P. A. Watters, and D. J. Tolhurst. A comparison of natural-image-based models of simple-cell coding. *Perception*, 29:1017–1040, 2000.

[25] D. J. Graham, D. M. Chandler, and D. J. Field. Can the theory of 'whitening' explain the center-surround properties of retinal ganglion cell receptive fields? *Vision Research*, 46:2901–2913, 2006.

[26] D. L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37:3358–3398, 1997.